GHENT UNIVERSITY

FACULTY OF SCIENCES

Marine Species Distributions: From data to predictive models

Samuel Bosch

Promoter: Prof. Dr. Olivier De Clerck

Thesis submitted in partial fulfilment of

the requirements for the degree of

Doctor (PhD) in Science – Biology

Academic year 2016-2017

VLIZ

UGent Phycology

Members of the examination committee

Prof. Dr. Olivier De Clerck - Ghent University (Promoter)*

Prof. Dr. Tom Moens – Ghent University (Chairman)

Prof. Dr. Elie Verleyen – Ghent University (Secretary)

Prof. Dr. Frederik Leliaert – Botanic Garden Meise / Ghent University

Dr. Tom Webb – University of Sheffield

Dr. Lennert Tyberghein - Vlaams Instituut voor de Zee

* non-voting members

Financial support

This thesis was funded by the ERANET INVASIVES project

(EU FP7 SEAS-ERA/INVASIVES SD/ER/010) and by

VLIZ as part of the Flemish contribution to the LifeWatch ESFRI.

Table of contents

Chapter 1

General Introduction

Species distribution modelling

Throughout most of human history knowledge of species diversity and their respective distributions was an essential skill for survival and civilization. In this respect it is not surprising that distributions of species and the mechanisms governing the distributions has intrigued humans from the early dawn of humanity up till present. Some of the earliest scientific writings about distributions of species and the interrelationships between species and the environment in which they live are found in *History of Animals* by Aristotle and *Enquiry into Plants* by Theophrastus (Mayr, 1985; Magner, 2002). More recently the link between species distributions, geography and the physical environment was noted by the likes of von Humboldt and Bonpland (1805), Watson (1847), de Candolle (1855), Wallace (1876) and Grinnell (1904), gradually leading to the advent of the research fields ecology and biogeography.

While these early writings were merely of a qualitative and descriptive nature, in the 21th century, ecology and biogeography under the influence of e.g. Hutchinson, Elton, McArthur and Wilson became progressively infused with theory and mathematics, thereby paving the way for species distribution modelling. The core of predictive modelling in geographic space involves the quantification of species-environment relationships (Guisan & Zimmermann, 2000). Species distribution modelling (SDM) as a separate research field emerged from the intersection of ecological gradient analysis (Whittaker et al., 1973), biogeography (Box, 1981), remote sensing and geographic information science (Franklin, 1995). SDM, also known as ecological niche modelling, habitat suitability modelling or climate envelope modelling, is now a widely used method in ecology, conservation biology, biogeography, paleoecology and global change biology. Apart from modelling species distributions, the same methods used in SDM have also been used in other fields to model for instance the geographical variation in species traits (Briones et al., 2014), agricultural crops (Hijmans et al., 2003), wildfire frequency (Syphard et al., 2008) and the distribution of humans and Neanderthals (Banks et al., 2008; Benito et al., 2017).

Species distribution modelling (SDM) is the process of using numerical tools to combine observations of species occurrence or abundance with environmental information (Elith & Leathwick, 2009). One of the core assumptions of SDM is that species are in equilibrium with their environment whereby a certain species occupies all suitable habitats (Austin, 2002; O'Connor, 2002; Araújo & Peterson, 2012). Invasive species in theory violate this assumption because they are expected

to be still expanding their range in the invaded region. These static, correlative models are opposed to more dynamic models of ecosystem processes (Guisan & Theurillat, 2000; Guisan & Zimmermann, 2000). Disturbances of the equilibrium between a species and its environment due to migration, invasion or biotic interactions put a natural upper limit on the performance of species distribution models. Some authors attempt to overcome this by building mechanistic distribution models, which incorporate information on the response of a species to different environmental conditions (Kearney & Porter, 2009). Others, combine both environmental and species co-occurrence data into a joint species distribution model (Clark et al., 2014; Harris, 2015; Ovaskainen et al., 2015; Tikhonov et al., 2017). Alternatively some studies include a temporal or dispersal aspect into their models (Zurell et al., 2009; Gutt et al., 2012; Génard & Lescourret, 2013; Mieszkowska et al., 2013; Hayes et al., 2015), resulting in models able to capture the dynamic aspect of the species distribution better. Such models, however, require additional knowledge on dispersal capacity and ecophysiology of the species that is often unavailable for marine species.

Regardless of the flavour of the species distribution model, all models require an estimate of the species distribution, either through presence-only or presence-absence data, as well as environmental data which is relevant for the geography of the species. Evidently, inadequate estimates of the distribution and environmental information, will reflect on the quality of the SDM.

Niche

The theoretical framework of SDM is based on the ecological niche concept (Guisan & Zimmermann, 2000; Pulliam, 2000). Biologists commonly distinguish three types of niches: the “Grinellian”, the “Eltonian” and the “Hutchinsonian”. Grinnell (1917) restricted the term niche to the climatic and habitat conditions where a species occurs (Pulliam, 2000; Guisan & Thuiller, 2005; Peterson et al., 2011). Elton (1927), on the other hand, viewed the niche as the functional role of a species in a community, especially its position in food webs, thereby highlighting species interactions. Lastly, Hutchinson (1957) defined the niche as “… the hypervolume defined by the environmental dimensions within which that species can survive and reproduce.” He furthermore distinguished the fundamental and realized niche, where the fundamental niche is the response of the species to the environment without taking species interactions into account and the realized niche takes species interactions into account. As correlative species distribution modelling starts from the realized distribution in geographic space it is only able to model the realized niche in environmental space (Austin, 2002; Colwell & Rangel, 2009). Translated to

geographic space, this realized niche may include areas where the species is not present due to dispersal limitation or biotic interactions. For species with source-sink populations even modelling the realized niche can be problematic when occurrences are recorded in unsuitable geographic and environmental space (Pulliam, 2000; Austin, 2002). For instance an invasive seaweed might be recorded shortly after its introduction in places where the year round conditions are not suitable for its long term survival and reproduction. But, if this record is used in a species distribution model it will lead to an overprediction of suitable areas.

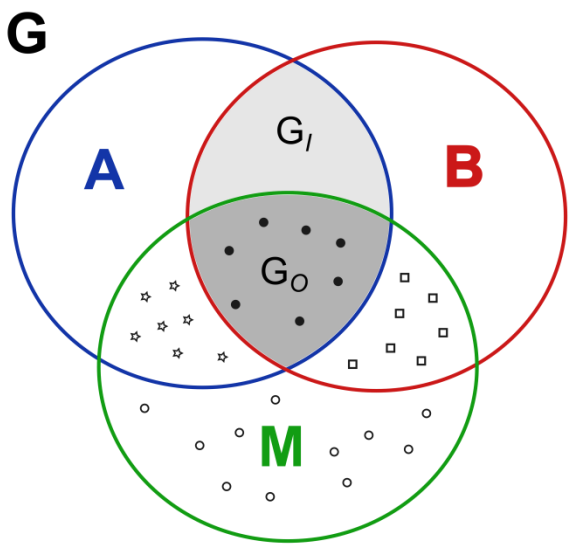


Figure 1. BAM diagram adapted from Soberón (2007) and Peterson et al. (2011) with the suitable abiotic (A) and biotic (B) conditions and the accessible area (M). Solid circles represent source populations, open shapes absences. The intersection between A and B is comprised of G_O, which is the occupied distributional area, and G_I, which is the potentially invadable distributional area. Note that the species absent in all areas except G_O, but only areas relevant for niche modelling have absences depicted. Open stars represent areas where the species is not competitive, open squares are areas with abiotically unsuitable conditions. G is the total area of the study.

Soberón and Peterson (2005) identified four factors determining the distribution of a species: abiotic conditions (A), biotic factors (B), regions that are accessible through dispersal (M) and the evolutionary capacity of the species to adapt to new conditions. Building on previous work by Pulliam (2000), Soberón and Peterson (2005) developed the BAM diagram, a Venn diagram representing all these distribution determining factors except evolution (Fig. 1). The center of the BAM diagram is the area occupied by the species (G_0), the invadable area, which is geographically not accessible through migration but abiotically and biotically suitable is denoted as G_I.

Distribution data

Most commonly used species distribution modelling algorithms require species occurrence and sometimes species absence data as modelling input. Ideally, such distribution data used for SDM comes from a well-designed biological survey, with accurately located species presences and absences from a known and uniform

sampling area (Franklin, 2009). However, an ever increasing amount of species occurrence information is available online, both from well-designed surveys and more anecdotal data from e.g. specimen collections and historical surveys. For example, the Global Biodiversity Information Facility (GBIF) currently hosts over 700 million distribution records from 1.6 million species. The largest marine database of species occurrences, the Ocean Biogeographic Information System (OBIS, Grassle, 2000), has grown from 400,000 records in 2002 to over 45 million records in 2017.

Unfortunately, not all data are of the same quality as they are subject to uneven sampling, taxonomic misidentification, errors in spatial coordinates and other data entry errors. Hence there is a need for adequate quality control of the provided data. Such errors have an even more significant impact when erroneous distribution points are located at the edge or outside of the true distribution of the species (Graham et al., 2007a; Naimi et al., 2014). In order to mitigate this issue Robertson et al. (2016) developed the R package *biogeo* for detecting and correcting errors in occurrence data and for the assessment of data quality of datasets from museum collections. In Chapter 2 we present a system implemented on the OBIS database that assigns various quality control (QC) flags to the records. These flags range from taxonomic and geographic issues, completeness of data to outlier detection. The OBIS R client, *robis* (Provoost et al., 2016), can be used to filter distribution records based on these QC flags.

The issue of sample selection bias is however not solved by quality control. Some of the proposed approaches to handle this involve filtering out species occurrences which are spatially (Veloz, 2009; Beck et al., 2014; Boria et al., 2014; Aiello-Lammens et al., 2015) or environmentally (Varela et al., 2014) close to each other. Additionally, for presence-only methods, a target group background can be used (Dudík et al., 2005; Phillips et al., 2009). With a target group background, background data is generated by randomly selecting species observations from a set of species which have been sampled in a similar way as the target species thus effectively creating background data with the same sample selection bias as the occurrence data used. Fourcade et al. (2014) concluded, for the presence-only algorithm MaxEnt, that systematically sub-sampling distribution records was generally the best performing methods for reducing bias in sample selection. For marine species, occurrence records are generally biased towards the coast and shallower waters (Robinson et al., 2011).

The best species distribution models can be obtained when models are fitted with reliable absences (Wisz & Guisan, 2009; Smith et al., 2013). But, these are rarely

reported for marine species and are currently not supported by public databases such as OBIS and GBIF. Absences can however in some cases be estimated for species collected in systematic surveys (Coro et al., 2016). As imperfect detection of marine species is very common, estimated absences are unreliable for all but the most surveyed and detectable species (Ready et al., 2010; Monk, 2013). Therefore, most applied marine SDM is presence-only modelling, where absences are substituted by a generated sample of observations that characterize the available environment, termed pseudo-absences or background data (Franklin, 2009).

Many studies on the selection of pseudo-absence or background data, and closely related to this, the extent of the study area, have been published (Phillips & Dudík, 2008; VanDerWal et al., 2009; Wisz & Guisan, 2009; Anderson & Raza, 2010; Lobo & Tognelli, 2011; Stokland et al., 2011; Barbet-Massin et al., 2012; Hanberry et al., 2012; Senay et al., 2013; Assis et al., 2014). Chefaoui & Lobo (2008) showed that the method of pseudo-absence selection has a direct influence on model predictions. They concluded that selecting pseudo-absence points randomly leads to species distribution models that predict smaller suitable areas and more closely resemble the realized distributions. Conversely, when pseudo-absence data is filtered in order to increase the environmental distance between presence and pseudo-absence data then model predictions reflect the potential distribution. In general there are two aspects to be considered with respect to pseudo-absence selection: the number of points and the location of points. For both aspects no consensus has been reached yet. Furthermore, the correct decision also depends on the goal of the study since the background data should reflect the environmental conditions and spatial extent of the ecological question of interest (Barve et al., 2011; Saupe et al., 2012).

Predictors

The predictors used for species distribution modelling generally consist of various environmental data sources such as climatological, topographic, geological and nutrient related maps (Franklin, 2009). However, the preparation of environmental data for SDM is a time-consuming task which has fuelled several independent initiatives to compile global datasets of environmental variables at a uniform spatial resolution and projection. Such datasets include WordClim for terrestrial species and Bio-ORACLE for marine species (Hijmans et al., 2005; Tyberghein et al., 2012). Do note that a significant number of studies compile their own data as not all data is available on a global scale or available at the desired resolution. In order to promote the use of these and other pre-compiled datasets we developed an R package *sdmpredictors*, presented in Chapter 3, which allows the exploration and

downloading of environmental data from WorldClim, ENVIREM, Bio-ORACLE and MARSPEC.

Environmental predictors of species distributions can be classified in two distinct ways: idealised types and distal or proximal predictors (Austin, 2002). The three idealised kinds of predictors are: direct, resource and indirect (Austin, 1980). Direct predictors have a direct physiological influence on the survival of a species but are not consumed by them (e.g. temperature). Resource predictors are consumed (e.g. light, nutrients) and indirect predictors have no physiological effect (e.g. longitude). The correlation between indirect predictors and the species distribution is due to their local correlation with one or more direct or resource predictors. Proximal and distal on the other hand refer to how close a predictor is in the chain of processes linking a predictor to its impact on the species. For example, the available sunlight at the surface of the blades of a seaweed would be a more proximal resource gradient than photosynthetically active radiation. The most robust and transferable species distribution models will be obtained when including only direct and proximal predictors (Austin, 2002). Moreover, as the shape of the species response to an indirect predictor depends on the nature of the correlation between the indirect and the direct and resource predictors, it can take any form (Guisan & Zimmermann, 2000). Indirect variables replace in many cases a combination of proximal direct and resource predictors.

Next to these abiotic factors shaping the distribution of species, some modellers include biological predictor variables in species distribution models. Accounting for biological factors such as the distribution of habitat forming species, dispersal range, species aggregations and interactions is likely to increase the performance of species distribution models, especially at landscape scale (Guisan et al., 2006; Nyström Sandman et al., 2013; Reiss et al., 2015).

Although the choice of the of environmental predictors has a large influence on SDM performance, the selection of predictors is not always obvious. This problem lead to several studies examining how to select the predictors that are to be used by the species distribution modelling algorithm. Brandt et al. (2017) noted that for broad-scale SDM, using statistical methods of variable selection is a useful first step. Statistical variable selection can for example be performed by inspecting the variable importance from random forests (Breiman, 2001; Genuer et al., 2015) or MaxEnt (Phillips et al., 2004) or by using Bayesian variable selection methods (O'Hara & Sillanpää, 2009). This statistical selection should be followed by a selection of predictors based on expert input in order to refine the models. These results

corroborate the findings by Petitpierre et al. (2017), who found that the a priori selection of ecologically meaningful predictors showed on average better transferability of SDM. On the other hand some other studies found that only using statistical methods of predictor selection results in adequate models (Pearce et al., 2001; Seoane et al., 2005; Charney, 2012). Alternatively, instead of a priori selecting predictors, Barbet-Massin and Jetz (2014), in their study on the relevance of different predictors for modelling bird distributions in North-America, assessed predictor relevance by comparing the results of models build with all possible combinations of predictors from different correlation groups. Predictors were grouped if members of a group had a Pearson correlation > 0.7 as including predictors with high collinearity leads to unstable parameter estimates (Dormann et al., 2013). Another approach is to remove correlations by using principal components analysis (PCA). PCA produces perfectly uncorrelated axes as output, providing an effective replacement for correlated variables (Dormann et al., 2013). Selecting the two first axes of a PCA calibrated on all predictors results on average in a better SDM transferability (Petitpierre et al., 2017). However, several disadvantages are linked to models fitted with principal components. Principal components are less ecologically interpretable and because correlations between predictors can change in the future they can't be used for climate change predictions (Janekovi & Novak, 2012; Petitpierre et al., 2017).

A cursory review of marine SDM studies published between 2003 and 2013 shows that 29 out of 49 studies mention biological reasons, expert opinion or previous studies as a justification for the included predictors, while 15 studies mention no reason and the remaining five studies selected the predictors based on availability and correlation. As for the number of predictors included, this ranged from 3 to 26 predictors with a median of 7 (see Supporting information for the list of studies).

Algorithms

Various algorithms or modelling methods are used for modelling species distributions (Franklin, 2009). Most commonly used methods have their origins in statistics or machine learning and differ in both how model are fitted and in the complexity of the resulting models. However differences in performance among different types of models tend to be smaller than differences among species (Elith et al., 2006; Franklin, 2009).

We distinguish two types of algorithms based on, whether or not they require (pseudo-)absence data. Presence-absence and presence-background methods are further divided into statistical and machine learning methods. Some examples of

methods using only occurrence records include envelope methods such as BIOCLIM (Busby, 1991) and environmental distance methods such as DOMAIN (Carpenter et al., 1993) or the Mahalanobis distance (Clark et al., 1993). BIOCLIM fits a multidimensional envelope model, which predicts the presence of species in environmental conditions that are not outside the 5% most extreme conditions, where the presence of a species has been previously reported.

The most commonly used statistical presence-absence methods are: generalized linear models (Nelder & Wedderburn, 1972), generalized additive models (Hastie & Tibshirani, 1986) and multivariate adaptive regression splines (Friedman, 1991). Furthermore, some Bayesian approaches have been applied to SDM (Franklin, 2009). Supervised machine learning methods based on presence-absence data, on the other hand, include decision trees based methods such as random forests (Breiman, 2001) and boosted regression trees (Friedman, 2001), artificial neural networks (McCulloch & Pitts, 1943; Hopfield, 1982), genetic algorithms (GARP, Stockwell & Noble, 1992) and support vector machines (Cortes & Vapnik, 1995).

Presence-absence methods can be applied to presence-only data if a sample of available locations, called pseudo-absence or background data, is generated appropriately (Franklin, 2009). Some methods like ecological niche factor analysis (ENFA, Hirzel et al., 2002), maximum entropy (MaxEnt, Phillips et al., 2004), expectation maximization (Ward, 2006), Poisson point process models (Warton & Shepherd, 2010) and maximum likelihood (Royle et al., 2012) have been specifically developed for this.

Recently two new SDM methods have been proposed. The first one, GRaF, uses a Bayesian machine learning technique called Gaussian random fields to create models (Golding & Purse, 2016). The second one, Plateau, attempts to create ecologically plausible climate envelopes by restricting the shape of the relationship between species distributions and climatic variables in spatial Bayesian species distribution models (Brewer et al., 2016).

Numerous studies have compared the performance of SDM algorithms (Elith et al., 2006; Guisan et al., 2007b; Meynard & Quinn, 2007; Tsoar et al., 2007; Ready et al., 2010; Lorena et al., 2011; Bucklin et al., 2015; García-Callejas & Araújo, 2015). From these we can conclude that more advanced methods like random forests, boosted regression trees and MaxEnt generally perform very well. But several authors noted that there is considerable variation between species and regions. An additional factor is that each algorithm has several settings that can be tuned, which tends to be a difficult task when modelling a large number species as is commonly done in

comparative studies. Indeed several studies have shown that species-specific tuning results in a significant improvement of the species distribution models by mitigating problems induced by sample selection bias and better transferability due to the selection of appropriate model complexities (Anderson & Gonzalez, 2011; Warren & Seifert, 2011; Radosavljevic & Anderson, 2014; Moreno-Amat et al., 2015). However, for rare species using a community-level approach for tuning instead of species-specific tuning is more appropriate (Madon et al., 2013).

The inconsistency in algorithm performance across species has led to the aggregation of results from different algorithms into an ensemble model (Araújo & New, 2007; Marmion et al., 2009; Buisson et al., 2010; Crimmins et al., 2013). The advantages of ensemble methods are that they can reduce the risk of choosing a wrong hypothesis or local minimum and it may be possible to expand the space of representable functions, and thus form a more accurate approximation to the true unknown hypothesis (Zhou, 2012). An important aspect of ensemble modelling is the ensemble diversity: the difference among the individual learners. The individual learners must be different in order to be able to improve the performance and they must not be very poor (Tumer & Ghosh, 1996; Zhou, 2012). While Marmion et al. (2009) found that the usage of ensemble models may significantly increase the accuracy of species distribution models, Crimmins et al. (2013) are in disfavour of ensemble models as they don't provide superior models while decreasing the ecological interpretability of the models.

Evaluation

In order to evaluate species distribution models there is a need for independent test data. Special attention has been given to the creation of cross-validation datasets, which is a non-trivial aspect given the spatial and sometimes temporal nature of the data (Arlot & Celisse, 2010; Roberts et al., 2016). An alternative or supplementary approach is to correct evaluation metrics based on the result of a null model (Raes & ter Steege, 2007; Hijmans, 2012).

The choice of evaluation metric is ideally made based on the goal of the study, whereby different weights are given to different errors (Guisan & Zimmermann, 2000; Mouton et al., 2010). Although criticised by Lobo et al. (2008), the area under the curve of the receiver operating characteristic (AUC, Hanley & McNeil, 1982) is the most commonly used evaluation metric for SDM. A large number of other metrics have been used, proposed and compared (Boyce et al., 2002; Allouche et al., 2006; Hirzel et al., 2006; Hand, 2009; Liu et al., 2011; Márcia Barbosa et al., 2013). Evaluation metrics can be distinguished based on two characteristics: whether they

need a threshold and whether they use test (pseudo-)absence data for evaluating the models. An example of a presence-only metric is the Boyce index (Boyce et al., 2002), which measures whether the number of predicted cells for a series of threshold values correlates with the number of evaluation points. The main appeal of the AUC is that it is a threshold independent metric. In order to calculate threshold dependent evaluation metrics, a method for picking the optimal threshold has to be selected. While various methods for doing this have been proposed, the technically optimal method is to select the threshold where the sum of the sensitivity (number of true positives / number of positive cases) and specificity (number of true negatives / number of negative cases) is maximized (Liu et al., 2013). Thresholding of suitability maps, and thus converting them into binary maps, is a common procedure as it allows for an easier interpretation and facilitates decision making. But, it is advised to also distribute the continuous probability maps, thus enabling the end user to select thresholds based on the specific objective of the study (Freeman & Moisen, 2008).

Forecasting

Several studies have identified anthropogenic climate change as one of the major threats to biodiversity, next to habitat destruction, pollution (eutrophication), invasive species and overexploitation of natural resources (Thuiller et al., 2005; Brook et al., 2008; Pereira et al., 2010). SDM is commonly used to predict species' range shifts under future climate scenarios (Hijmans & Graham, 2006; Jueterbock et al., 2013; Pearson et al., 2013). But careful interpretation of these species distribution models is needed as the predictive accuracy can be poor (Rapacciuolo et al., 2012; Smith, 2013).

All aspects of SDM (occurrences, background, predictors, algorithms and model selection) have an impact on future climate predictions of species distributions. Both Synes and Osborne (2011) and Braunisch et al. (2013) noted that selection of predictors has a particularly high impact. Moreover the number of predictors available for future climate predictions is limited. For future climate change modelling additional uncertainty is introduced due to the availability of different global circulation models and climate change scenarios (Buisson et al., 2010). Related to this, Stoklosa et al. (2015), showed that errors in environmental data can lead to biased coefficient estimates in the species distribution models and proposed a framework for integrating this uncertainty by creating maps depicting uncertainty.

Furthermore, as the distribution of species is also shaped by biotic interactions and dispersal limitations, changes in these biotic interactions with changing

environmental conditions under climate change may lead to considerable additional uncertainty in their future distribution (Pearman et al., 2008; Robinson et al., 2011). Recent advances in joint species distribution modelling aim to uncover and integrate these biotic interactions in SDM in order to improve predictions for the current and future climate (e.g. Clark et al., 2014; Pollock et al., 2014). Moreover, even without changes in biotic interactions, predicting the future distribution assumes the absence of niche shifts which is not always the case (Pearman et al., 2008; Early & Sax, 2014; Guisan et al., 2014).

Seaweeds as a case study

Some application areas for marine SDM include marine spatial planning, the creation of monitoring designs, assessing the risks involved with non-native species and predicting future distributions in order to account for climate change (Reiss et al., 2015). However, some marked differences between the terrestrial and marine modelling setup are obvious. Firstly, sampling is more expensive in the marine environment, resulting in occurrence records with a lower spatial resolution and which are biased towards the coast and economically important areas (Robinson et al., 2011; Reiss et al., 2015). Moreover, the detectability of many marine species is much lower than for terrestrial species (MacLeod et al., 2008). Additionally the environmental data is equally less sampled with a higher reliance on remotely sensed data and fewer in situ data points as compared to the terrestrial environment, which limits the resolution and accuracy of the available data.

Next to modelling a broad range of marine species in Chapter 4, we selected seaweeds as a case study for marine SDM as their distribution is strongly affected by environmental factors (Lüning, 1990; Adey & Steneck, 2001). More specifically its global distribution is mainly limited by temperature, while other abiotic factors, such as bathymetry, substrate type and available light, play a role at a regional or local scale (Lüning, 1990). However, seaweed species used for modelling should be selected with care as they are prone to misidentifications (Marcelino & Verbruggen, 2015). Earlier SDM studies have used seaweeds to demonstrate the usability of the Bio-ORACLE dataset (Tyberghein et al., 2012), to model invasive seaweeds (Verbruggen et al., 2013) and to assess the impact of future climate change (Jueterbock et al., 2013; Assis et al., 2014; Martínez et al., 2015).

Introduced seaweeds

The study of introduced species is an important part of ecology as they are considered to be a major threat for native species communities (Norse, 1993;

Molnar et al., 2008; Winter et al., 2009). The introduction of species systematically results in biotic homogenization and changes in ecosystem functioning (Olden et al., 2004; Hooper et al., 2005; Sousa et al., 2009; Winter et al., 2009). Europe is a hot-spot for aquatic introductions with around 600 alien species established at present (Gollasch, 2006). In this thesis we define alien species as species that are introduced outside their natural geographic range due to human activities, while, invasive species are introduced species that are spreading at such a rate that they are damaging the environment, economy or human health.

Alien seaweeds represent one of the largest groups of marine aliens in Europe, and constitute between 20 and 29 % of all alien marine species (Schaffelke et al., 2006). Seaweeds are major primary producers in coastal areas, and are extremely important for coastal ecosystems by supporting high biodiversity through structuring complex habitats for associated species. Large-scale substitution of dominant native seaweeds with alien species will consequently alter coastal productivity and food web structure, and therefore impact ecosystem services. Only a few impact studies on invasive seaweeds have been carried out worldwide, and these have detected a range of negative ecological effects, with reduction in abundance of native biota being most frequently reported (Williams & Smith, 2007). Rising temperatures will most likely impact the alien fauna component more than the native one and cause increasing abundances of the alien component (Sorte et al., 2010), but very little is known about how temperature variation influences the relationship between alien and native seaweeds. But, previous studies have demonstrated that temperature is a key parameter for the distribution of some invasive seaweeds (Nejrup et al., 2013; Samperio-Ramos et al., 2015; Cecere et al., 2016).

Species are introduced unintentionally (e.g. shipping) or intentionally (e.g. aquaculture) (Gollasch, 2007). Boat traffic and aquaculture, in particular oyster import from Asia, have been identified as the most important vectors for introduced seaweeds in Europe (Mineur et al., 2008, 2014). It is however likely that other significant vectors exist. A more complete understanding of the introduction process is paramount for predicting future spread. Prediction of the future range of invasive seaweeds is important for risk assessment and future management. Establishment of invasive seaweeds is strongly linked to environmental conditions. Previous risk analyses have tried to match traits of alien seaweeds to environmental factors (Nyberg & Wallentinus, 2005), but these analyses have not been able to include climate variation as a factor. Advanced species distribution modelling techniques present a more powerful way to predict range extensions and shifts and allow

predicting the range of invasive seaweeds in future climatic scenarios (Tyberghein et al., 2012; Verbruggen et al., 2013).

The predictive performance of distribution models of introduced species depends on the degree of niche conservatism between the native and invaded range (Pearman et al., 2008). For introduced species we distinguish two types of niche shifts: 1) niche shifts into environmental conditions in the invaded range that are available in the native range (analog conditions) and 2) niche shifts into novel conditions (non-analog conditions) (Guisan et al., 2014). While niche conservatism in analog conditions has been shown to be prevalent for introduced terrestrial plants and birds (Petitpierre et al., 2012; Strubbe et al., 2013), this has not yet been confirmed for introduced seaweeds.

Aims and outline

The overall objective of this thesis is to improve and contribute to the process and understanding of marine species distribution modelling in order to facilitate an in depth study of the trends, vectors and distribution of introduced seaweeds in Europe.

In Chapter 2 we provide quality indicators for the marine species distribution data available in the European and international Ocean Biogeographic Information Systems (EurOBIS and OBIS). Next to various checks on data integrity and completeness, outliers in geographic and environmental space were identified, allowing end)users to select distribution data that is fit for their specific purposes.

In Chapter 3 we make global environmental datasets for species distribution modelling in the past, current and future climate more accessible. This is achieved by developing an R package that facilitates the usage of various published marine and terrestrial environmental datasets for species distribution modelling.

Based on the results from chapters 2 and 3 we developed a marine benchmark dataset with distribution data and environmental data for more than 500 species (MarineSPEED). With this dataset we aim to get a better understanding of the relevance of different predictors for modelling the distribution of marine species for a broad range of modelling setups.

While the first three chapters concerned general marine species distributions, the next three chapters explore the specific case of introduced seaweeds in Europe. In

Chapter 5 we aim to analyse and map the introduction history, origin and trends in introduced seaweeds in Europe.

In Chapter 6 we evaluate the risk of aquaria and aquarium trade as a vector for future introductions of seaweeds in Europe. After assessing the seaweed diversity on e-commerce websites and in local aquaria we mapped the current and future ecoregions in Europe that are potentially suitable for non-native seaweeds available in aquaria based on thermal niche models.

Chapter 7 aims to build species distribution models that are able to predict the future distribution of introduced seaweeds before and after their introduction in Europe. We additionally set out to elucidate differences in model performance for the modelled introduced seaweeds by measuring the niche expansion. Additionally, we propose a method for identifying candidate areas for further spreading under global climate change.

Finally, in the general discussion we highlight common aspects of the different chapters in this thesis such as data and uncertainty. Furthermore, we discuss future research avenues.

Supporting information

List of marine species distribution modelling studies, published between 2003 and 2013, that were reviewed in order to record the used justifications for the selected predictors and to record the number of predictors used:

Acevedo P., Jiménez-Valverde A., Lobo J.M., & Real R. (2012) Delimiting the geographical background in species distribution modelling. *Journal of Biogeography*, **39**, 1383–1390.

Bailey H. & Thompson P. (2009) Using marine mammal habitat modelling to identify priority conservation zones within a marine protected area. *Marine Ecology Progress Series*, **378**, 279–287.

Beaugrand G., Lenoir S., Ibañez F., & Manté C. (2011) A new model to assess the probability of occurrence of a species, based on presence-only data. *Marine Ecology Progress Series*, **424**, 175–190.

Beaugrand G., Mackas D., & Goberville E. (2013) Applying the concept of the ecological niche and a macroecological approach to understand how climate influences zooplankton: Advantages, assumptions, limitations and requirements. *Progress in Oceanography*, **111**, 75–90.

Bentlage B., Peterson A.T., Barve N., & Cartwright P. (2013) Plumbing the depths: extending ecological niche modelling and species distribution modelling in three dimensions. *Global Ecology and Biogeography*, **22**, 952–961.

Compton T.J., Leathwick J.R., & Inglis G.J. (2010) Thermogeography predicts the potential global range of the invasive European green crab (*Carcinus maenas*). *Diversity and Distributions*, **16**, 243–255.

Downie A.-L., von Numers M., & Boström C. (2013) Influence of model selection on the predicted distribution of the seagrass *Zostera marina*. *Estuarine, Coastal and Shelf Science*, **121–122**, 8–19.

Ellis J., Ysebaert T., Hume T., Norkko A., Bult T., Herman P., Thrush S., & Oldman J. (2006) Predicting macrofaunal species distributions in estuarine gradients using logistic regression and classification systems. *Marine Ecology Progress Series*, **316**, 69–83.

Elsäßer B., Fariñas-Franco J.M., Wilson C.D., Kregting L., & Roberts D. (2013) Identifying optimal sites for natural recovery and restoration of impacted biogenic habitats in a special area of conservation using hydrodynamic and habitat suitability modelling. *Journal of Sea Research*, **77**, 11–21.

Embling C.B., Gillibrand P. a., Gordon J., Shrimpton J., Stevick P.T., & Hammond P.S. (2010) Using habitat models to identify suitable sites for marine protected areas for harbour porpoises (*Phocoena phocoena*). *Biological Conservation*, **143**, 267–279.

Freitas C., Kovacs K.M., Ims R. A., & Lydersen C. (2008) Predicting habitat use by ringed seals (*Phoca hispida*) in a warming Arctic. *Ecological Modelling*, **217**, 19–32.

Friedlaender A.S., Johnston D.W., Fraser W.R., Burns J., Patrick N. H., & Costa D.P. (2011) Ecological niche modeling of sympatric krill predators around Marguerite Bay, Western Antarctic Peninsula. *Deep Sea Research Part II: Topical Studies in Oceanography*, **58**, 1729–1740.

Gogina M., Glockzin M., & Zettler M.L. (2010) Distribution of benthic macrofaunal communities in the western Baltic Sea with regard to near-bottom environmental parameters. 2. Modelling and prediction. *Journal of Marine Systems*, **80**, 57–70.

Gogina M. & Zettler M.L. (2010) Diversity and distribution of benthic macrofauna in the Baltic Sea. *Journal of Sea Research*, **64**, 313–321.

Hattab T., Ben Rais Lasram F., Albouy C., Sammari C., Romdhane M.S., Cury P., Leprieur F., & Le Loc'h F. (2013) The Use of a Predictive Habitat Model and a Fuzzy Logic Approach for Marine Management and Planning. *PLoS ONE*, **8**, e76430.

Howell K.L., Holt R., Endrino I.P., & Stewart H. (2011) When the species is also a habitat: Comparing the predictively modelled distributions of *Lophelia pertusa* and the reef habitat it forms. *Biological Conservation*, **144**, 2656–2665.

Huang Z., Brooke B., & Li J. (2011) Performance of predictive models in marine benthic environments based on predictions of sponge distribution on the Australian continental shelf. *Ecological Informatics*, **6**, 205–216.

Huff D.D., Lindley S.T., Wells B.K., & Chai F. (2012) Green sturgeon distribution in the Pacific Ocean

estimated from modeled oceanographic features and migration behavior. *PloS one*, **7**, e45852.

Iken K., Konar B., Benedetti-Cecchi L., Cruz-Motta J.J., Knowlton A., Pohle G., Mead A., Miloslavich P., Wong M., Trott T., Mieszkowska N., Riosmena-Rodriguez R., Airoldi L., Kimani E., Shirayama Y., Fraschetti S., Ortiz-Touzet M., & Silva A. (2010) Large-scale spatial distribution patterns of echinoderms in nearshore rocky habitats. *PloS one*, **5**, e13845.

Inglis G.J., Hurren H., Oldman J., & Haskew R. (2006) Using habitat suitability index and particle dispersion models for early detection of marine invaders. *Ecological Applications*, **16**, 1377–1390.

Irwin A.J., Nelles A.M., & Finkel Z. V. (2012) Phytoplankton niches estimated from field data. *Limnology and Oceanography*, **57**, 787–797.

Jones M.C., Dye S.R., Pinnegar J.K., Warren R., & Cheung W.W.L. (2012) Modelling commercial fish distributions: Prediction and assessment using different approaches. *Ecological Modelling*, **225**, 133–145.

Jueterbock A., Tyberghein L., Verbruggen H., Coyer J. A., Olsen J.L., & Hoarau G. (2013) Climate change impact on seaweed meadow distribution in the North Atlantic rocky intertidal. *Ecology and Evolution*, **3**, 1356–1373.

Knudby A., Kenchington E., & Murillo F.J. (2013) Modeling the Distribution of Geodia Sponges and Sponge Grounds in the Northwest Atlantic. *PLoS ONE*, **8**, e82306.

Lassalle G., Crouzet P., Gessner J., & Rochard E. (2010) Global warming impacts and conservation responses for the critically endangered European Atlantic sturgeon. *Biological Conservation*, **143**, 2441–2452.

Leathwick J., Moilanen A., Francis M., Elith J., Taylor P., Julian K., Hastie T., & Duffy C. (2008) Novel methods for the design and evaluation of marine protected areas in offshore waters. *Conservation Letters*, **1**, 91–102.

Lenoir S., Beaugrand G., & Lecuyer É. (2011) Modelled spatial distribution of marine fish and projected modifications in the North Atlantic Ocean. *Global Change Biology*, **17**, 115–129.

MacLeod C.D., Mandleberg L., Schweder C., Bannon S.M., & Pierce G.J. (2008) A comparison of approaches for modelling the occurrence of marine animals. *Hydrobiologia*, **612**, 21–32.

Magris R. & Déstro G. (2010) Predictive modeling of suitable habitats for threatened marine invertebrates and implications for conservation assessment in Brazil. *Brazilian Journal of Oceanography*, **58**, 57–68.

Maravelias C., Haralabous J., & Papaconstantinou C. (2003) Predicting demersal fish species distributions in the Mediterranean Sea using artificial neural networks. *Marine Ecology Progress Series*, **255**, 249–258.

Maxwell D.L., Stelzenmüller V., Eastwood P.D., & Rogers S.I. (2009) Modelling the spatial distribution of plaice (*Pleuronectes platessa*), sole (*Solea solea*) and thornback ray (*Raja clavata*) in UK waters for marine management and planning. *Journal of Sea Research*, **61**, 258–267.

Merckx B., Steyaert M., Vanreusel A., Vincx M., & Vanaverbeke J. (2011) Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. *Ecological Modelling*, **222**, 588–597.

Monk J., Ierodiaconou D., Harvey E., Rattray A., & Versace V.L. (2012) Are we predicting the actual or apparent distribution of temperate marine fishes? *PloS one*, **7**, e34558.

Moura A.E., Sillero N., & Rodrigues A. (2012) Common dolphin (*Delphinus delphis*) habitat preferences using data from two platforms of opportunity. *Acta Oecologica*, **38**, 24–32.

Nyström Sandman A., Wikström S. a., Blomqvist M., Kautsky H., & Isaeus M. (2013) Scale-dependent influence of environmental variables on species distribution: a case study on five coastal benthic species in the Baltic Sea. *Ecography*, **36**, 354–363.

Oswald S. a., Huntley B., Collingham Y.C., Russell D.J.F., Anderson B.J., Arnold J.M., Furness R.W., & Hamer K.C. (2011) Physiological effects of climate on distributions of endothermic species. *Journal of Biogeography*, **38**, 430–438.

Palialexis A., Georgakarakos S., Karakassis I., Lika K., & Valavanis V.D. (2011) Prediction of marine species distribution from presence–absence acoustic data: comparing the fitting efficiency and the predictive capacity of conventional and novel distribution models. *Hydrobiologia*, **670**, 241–266.

Panigada S., Zanardelli M., MacKenzie M., Donovan C., Mélin F., & Hammond P.S. (2008) Modelling

habitat preferences for fin whales and striped dolphins in the Pelagos Sanctuary (Western Mediterranean Sea) with physiographic and remote sensing variables. *Remote Sensing of Environment*, **112**, 3400–3412.

Pittman S.J. & Brown K. a (2011) Multi-scale approach for predicting fish species distributions across coral reef seascapes. *PloS one*, **6**, e20583.

Praca E., Gannier A., Das K., & Laran S. (2009) Modelling the habitat suitability of cetaceans: Example of the sperm whale in the northwestern Mediterranean Sea. *Deep Sea Research Part I: Oceanographic Research Papers*, **56**, 648–657.

Ready J., Kaschner K., South A.B., Eastwood P.D., Rees T., Rius J., Agbayani E., Kullander S., & Froese R. (2010) Predicting the distributions of marine organisms at the global scale. *Ecological Modelling*, **221**, 467–478.

Reiss H., Cunze S., König K., Neumann H., & Kröncke I. (2011) Species distribution modelling of marine benthos: a North Sea case study. *Marine Ecology Progress Series*, **442**, 71–86.

de Rivera C.E., Steves B.P., Fofonoff P.W., Hines A.H., & Ruiz G.M. (2011) Potential for high-latitude marine invasions along western North America. *Diversity and Distributions*, **17**, 1198–1209.

Šiaulys A. & Bučas M. (2012) Species distribution modelling of benthic invertebrates in the south-eastern Baltic Sea. *Baltica*, **25**, 163–170.

Siders Z. a., Westgate A.J., Johnston D.W., Murison L.D., & Koopman H.N. (2013) Seasonal Variation in the Spatial Distribution of Basking Sharks (*Cetorhinus maximus*) in the Lower Bay of Fundy, Canada. *PLoS ONE*, **8**, e82074.

Torres L., Read A., & Halpin P. (2008) Fine-scale habitat modeling of a top marine predator: do prey data improve predictive capacity. *Ecological Applications*, **18**, 1702–1717.

Valle M., van Katwijk M.M., de Jong D.J., Bouma T.J., Schipper A.M., Chust G., Benito B.M., Garmendia J.M., & Borja Á. (2013) Comparing the performance of species distribution models of *Zostera marina*: Implications for conservation. *Journal of Sea Research*, **83**, 56–64.

Weinmann A.E., Rödder D., Lötters S., & Langer M.R. (2013) Heading for New Shores: Projecting Marine Distribution Ranges of Selected Larger Foraminifera. *PLoS ONE*, **8**, e62182.

Yesson C., Taylor M.L., Tittensor D.P., Davies A.J., Guinotte J., Baco A., Black J., Hall-Spencer J.M., & Rogers A.D. (2012) Global habitat suitability of cold-water octocorals. *Journal of Biogeography*, **39**, 1278–1292.

Chapter 2

Fishing for data and sorting the catch: assessing the data quality, completeness and fitness for use of data in marine biogeographic databases

Leen Vandepitte[1], Samuel Bosch[1,2], Lennert Tyberghein[1], Filip Waumans[1], Bart Vanhoorne[1], Francisco Hernandez[1], Olivier De Clerck[2] and Jan Mees[1]

[1]*Flanders Marine Institute (VLIZ), InnovOcean site, Wandelaarskaai 7, 8400 Ostend, Belgium*
[2]*Research Group Phycology, Biology Department, Ghent University, Krijgslaan 281/S8, 9000 Ghent, Belgium*

Published in January 2015 (doi: 10.1093/database/bau125).

SB implemented the EurOBIS quality control procedures on OBIS and developed and documented the outlier detection procedures.

Abstract

Being able to assess the quality and level of completeness of data has become indispensable in marine biodiversity research, especially when dealing with large databases that typically compile data from a variety of sources. Very few integrated databases offer quality flags on the level of the individual record, making it hard for users to easily extract the data that are fit for their specific purposes. This article describes the different steps that were developed to analyse the quality and completeness of the distribution records within the European and international Ocean Biogeographic Information Systems (EurOBIS and OBIS). Records are checked on data format, completeness and validity of information, quality and detail of the used taxonomy and geographic indications and whether or not the record is an outlier. The corresponding quality control (QC) flags will not only help users with their data selection, they will also help the data management team and the data custodians to identify possible gaps and errors in the submitted data, providing scope to improve data quality. The results of these quality control procedures are as of now available on both the EurOBIS and OBIS databases. Through the Biology portal of the European Marine Observation and Data Network (EMODnet Biology), a subset of EurOBIS records—passing a specific combination of these QC steps—is offered to the users. In the future, EMODnet Biology will offer a wide range of filter options through its portal, allowing users to make specific selections themselves. Through LifeWatch, users can already upload their own data and check them against a selection of the here described quality control procedures.

Database URL: http://www.eurobis.org (http://www.iobis.org; www.emodnet-biology.eu)

Introduction

Progress in information technology has resulted in an increasing flood of data and information. Efficiently mining this sea of data and determining the quality of the data and its fitness for use has become a major challenge of many disciplines. Evaluating and documenting the quality of data has already become a standard practice in several scientific disciplines over many years, e.g. in medicine (Congalton, 1991; Sherwood, 1991; Lunetta & Lyon, 2004; Garaba et al., 2011), remote sensing (Beissbarth et al., 2000; Pruesse et al., 2007; Otto et al., 2008) and gene sequencing (Chapman, 2005; Hill et al., 2010; Vandepitte et al., 2011). It is however only in the last decade that its importance—in combination with the assessment of the fitness for use—has become evident for biological sciences, more specifically for biodiversity data and data related to species occurrences (Yesson et al., 2007; Robertson, 2008; Vandepitte et al., 2010; Appeltans et al., 2012; Candela et al., 2015).

Biodiversity is inextricably linked with biogeography (Ray, 1996), which is clear from the many papers that contain both biodiversity and biogeography in their titles, abstracts and keywords (e.g. Wulff et al. 2009, O'Dor et al. 2010, Obura 2012, Selama et al. 2013). And both concepts are not only essential in research hypotheses, but also in the field of conservation, management (Ray, 1996; Richardson & Whittaker, 2010; Chiarucci et al., 2011) and modelling (Woolley et al., 2013; Bocedi et al., 2014; Convey et al., 2014).

When looking at larger patterns—e.g. on a European or global scale—data are mostly aggregated from a variety of sources. For the marine environment, data on all living marine species from different regional data centres and nodes flow towards the international Ocean Biogeographic Information System (OBIS; www.iobis.org), making marine biogeographic data freely available online. A variety of data is captured, going from data collected during research and monitoring campaigns to data from museum collections or data derived from literature. Given this very diverse nature of data, there is a strong need to be able to assess the quality of these data and provide feedback to the data providers. In addition, a system to assess the completeness of the record needed to be developed, offering specific filters to the users to be able to e.g. only query species records where complete abundance information is available.

Assessing the quality of a distribution record has thus become indispensable, as has the ability to give an indication of the completeness of that record, especially in database infrastructures such as e.g. EurOBIS, OBIS and the Global Biodiversity

Information Facility (GBIF; www.gbif.org) that provide access to data from a wide range of sources (e.g. Yesson et al. 2007, Robertson 2008). Several actions regarding quality control and data cleaning have already been undertaken on regional or group-specific databases such as SpeciesLink (http://splink.cria.org.br) for Brazilian data collections, Fauna Europaea (de Jong et al., 2014) for European land and freshwater animal species, fish collection databases in relation to FishBase (Froese et al., 1999) and the Atlas of Living Australia (ALA, http://www.ala.org.au/). However, efforts on quality control and fitness for use for marine biogeographic data were not yet globally organized, as is now presented here for OBIS.

An indication of the completeness can help the user in evaluating whether a particular record is useful for their analysis or not. A distribution record without a timestamp can e.g. be used to get insights in the general distribution of a species but will not be useful for temporal analysis. This illustrates that distribution records, although they do not share the same level of completeness, can be used for a multitude of applications, depending on the user's needs.

Over the last year, quality control (QC) tools have been developed to be able to document both the quality and completeness of each distribution record within EurOBIS. After extensive testing, these QC tools have been implemented in OBIS and extended with extra quality control procedures. This article will elaborate on these recently developed automated quality control procedures and their relevance. In addition, we will demonstrate the importance and usability of these procedures with some use cases. The main goal of these QC steps is to provide a measure of fitness for use of marine biogeographic data both for the scientists and data managers, by offering several tools that help assessing the completeness and validity of distribution records. For a general description of the structure and content of the EurOBIS and OBIS database, we respectively refer to (Grassle, 2000; Zhang & Grassle, 2002; Vandepitte et al., 2011).

Data systems

The quality control procedures were originally developed on EurOBIS, to add quality flags to the available data. Because these data are largely limited to European seas—and a number of QC steps only make sense on a global level (e.g. outlier detection)—the exercise was repeated on the OBIS database, with addition of a number of steps related to outlier analyses.

The QC procedures on EurOBIS were developed in two different ways: (1) as an automated process, to be able to assess the quality and completeness of the records already available within the database and (2) as online web services that can be used by potential data providers and researchers to assess the quality and completeness of their own data prior to use or submission. The former allows data managers to provide feedback to data providers and to check whether they can make their data more complete and correct gaps and putative errors. In addition, the results of the QC steps can be used for specific filtering on the data. The latter return a result report, listing all records that do not comply with a certain QC step. Users can immediately adapt their data and rerun the QC procedures online before analysing or submitting the data to EurOBIS.

EurOBIS is one of the many regional nodes within OBIS and is committed to a continuous support of OBIS, translated in serving its distribution data to OBIS. As the QC procedures also run on OBIS, the results of this can provide a valuable feedback to the other involved nodes and will therefore improve the quality and completeness of the online available records. Both the data providers and the separate nodes would benefit from this. From OBIS, data are sent to the Global Biodiversity Information Facility (GBIF), which would thus imply that GBIF could also only offer marine data that comply with a certain quality standard.

Quality control procedures

The quality control procedures have been developed for two main reasons. First of all, the available tools offer scientists the opportunity to quality check their data, prior to planned analyses or publishing their data through (Eur)OBIS and they help the (Eur)OBIS data management team in assessing the completeness and quality of the data when making them available online. When incomplete or possibly incorrect data are sent to (Eur)OBIS, the data management team can easily communicate with the provider on the possibly incorrect records based on the assigned quality flags. Secondly, the assigned quality flags can (i) help users in selecting data that are fit for their specific use and purpose or (ii) make it possible to filter records that comply with a certain quality standard and send those to other data systems such as e.g. the European Marine Observation and Data Network (EMODnet).

Each distribution record goes through a series of automated quality control steps, each generating a QC flag. Each QC step is a question that has a yes/no (= 1/0) answer and the result is stored as a bit-sequence ($2^{(x-1)}$) where x represents the number of the QC flag. The results of all these QC steps are added up and stored in a

single QC field in the (Eur)OBIS database, generating a unique integer value for each possible combination of positively evaluated QC steps. An overview of all the QC steps and their corresponding bit-sequence is given in Table 1. Given the different structure and scope of EurOBIS and OBIS, a number of QC steps have been specifically developed for either EurOBIS or OBIS. The majority (17) of the QC steps are, however, available for both data systems.

The strength of the quality control procedures is that they not only evaluate a dataset as a whole but also look at each record individually, giving a much more detailed view on the quality and completeness of the data and providing more opportunities to users in their data selection as one dataset may contain several useful records, which might have been rejected if the evaluation had been done solely on the dataset level.

1. Data format checks

Data made available through (Eur)OBIS need to be compliant with the OBIS Schema, used by OBIS. This OBIS Schema has 74 data and information fields, of which 7 are mandatory and 15 are highly recommended. The remaining fields are classified as optional. For a full overview of the OBIS Schema, we refer to the OBIS website (http://www.iobis.org/node/304). A lot of data providers are making use of the Integrated Publishing Toolkit (IPT) developed by GBIF (Robertson et al., 2014) to exchange their data. By doing so, their data follow the Darwin Core format (Wieczorek et al., 2012) which slightly differs from the OBIS Schema, which is based on an older version of the Darwin Core format. To avoid confusion, the EurOBIS website includes a mapping between the OBIS Schema field names and the currently used Darwin Core field names (http://www.eurobis.org/data_formats).

The data format check compares the general format of a dataset with the requirements of the OBIS Schema. When any of the required fields is missing or original field names are not correctly mapped to the field names used within OBIS, then these records are negatively evaluated in the QC procedures and are thus in need of an additional check. Fields that are not part of the OBIS Schema can still be shared with EurOBIS—e.g. through the DarwinCore Archive format (GBIF, 2011)—but the corresponding data will—at this time—not be shown through the data portal. If the OBIS Schema recommends the use of certain wording or codes—e.g. in the field ‘BasisOfRecord’—this is also checked. The ‘BasisOfRecord’ defines the kind of data: which can be actual observations (O), specimen information from museum collections (S) or distribution data derived from literature (L), which can already provide a first important data filter for the user.

Table 1. Overview of all the QC steps in the EurOBIS database, including the unique bit-sequence ($2^{(x-1)}$, with x = number of the QC flag) when the QC step is evaluated positively. The second last column lists whether a QC step is also available to the users through the online web services. IQR = Interquartile range; MAD = Median absolute deviation; SSS = Sea surface salinity; SST = Sea surface temperature.

QC	Category	Question	Bit-sequence, if answer is yes	Available as online data service	Implemented in
2	Taxonomy	Is the taxon name matched to WoRMS?	2	Yes (taxon match)	EurOBIS + OBIS
3	Taxonomy	Is the taxon level lower than family?	4	Yes (taxon match)	EurOBIS + OBIS
4	Geography: lat/lon	Are the latitude/longitude values different from zero?	8	Yes (check OBIS format)	EurOBIS + OBIS
5	Geography: lat/lon	Are the latitude/longitude values within their possible boundaries?	16	Yes (check OBIS format)	EurOBIS + OBIS
6	Geography: lat/lon	Are the coordinates situated in sea or along the coastline (20 km buffer)?	32	Yes (check OBIS format)	EurOBIS + OBIS
9	Geography: lat/lon	Are the coordinates situated in the expected geographic area (compare metadata)?	256	No, but visual check possible through separate data validation service	EurOBIS
18	Geography: depth	Is minimum depth ≤ maximum depth?	131 072	Not yet available	EurOBIS + OBIS
19	Geography: depth	Is the sampling depth possible when compared with GEBCO depth map (incl. margin)?	262 144	No, but depths per lat-lon can be requested through geographic web services	EurOBIS + OBIS
7	Completeness: date/time	Is the sampling year (start/end) completed and valid?	64	Yes (check OBIS format)	EurOBIS + OBIS
11	Completeness: date/time	Is the sampling date (year/month/day; start/end) valid?	1 024	Yes (check OBIS format)	EurOBIS + OBIS
12	Completeness: date/time	If a start and end date are given, is the start before the end?	2 048	Yes (check OBIS format)	EurOBIS + OBIS
13	Completeness: date/time	If a sampling time is given, is this valid and is the time zone completed?	4 096	Not yet available	EurOBIS + OBIS
14	Completeness: presence/abundance/biomass	Is the value of the field 'ObservedIndividualCount' empty or > 0?	8 192	Not yet available	EurOBIS + OBIS
15	Completeness:	Is the value of the field 'Observedweight' empty	16 384	Not yet available	EurOBIS + OBIS

QC	Category	Question	Bit-sequence, if answer is yes	Available as online data service	Implemented in
	presence/abundance/biomass	or > 0?			
16	Completeness: presence/abundance/biomass	Is the field 'SampleSize' completed if the field 'ObservedIndividualCount' is > 0?	32 768	Not yet available	EurOBIS + OBIS
1	(Eur)OBIS data format	Are the required fields from the OBIS Schema completed?	1	Yes (check OBIS format)	EurOBIS + OBIS
10	(Eur)OBIS data format	Is the 'Basis of Record' documented, and is an existing OBIS code used?	512	Yes (check OBIS format)	EurOBIS + OBIS
17	(Eur)OBIS data format	Is the value of the field 'Sex' empty or is an existing OBIS code used?	65 536	Not yet available	EurOBIS + OBIS
21	Outliers:environment	Is the observation within six MADs from the median depth of this taxon?	1 048 576	Not yet available	OBIS
22	Outliers:environment	Is the observation within three IQRs from the first & third quartile depth of this taxon?	2 097 152	Not yet available	OBIS
23	Outliers:environment	Is the observation within six MADs from the median SSS of this taxon?	4 194 304	Not yet available	OBIS
24	Outliers:environment	Is the observation within three IQRs from the first & third quartile SSS of this taxon?	8 388 608	Not yet available	OBIS
25	Outliers:environment	Is the observation within six MADs from the median SST of this taxon?	16 777 216	Not yet available	OBIS
26	Outliers:environment	Is the observation within three IQRs from the first & third quartile SST of this taxon?	33 554 432	Not yet available	OBIS
27	Outliers:geography	Is the observation within six MADs from the distance to the centroid of this taxon?	67 108 864	Not yet available	OBIS
28	Outliers:geography	Is the observation within three IQRs from the first & third quartile distance to the centroid of this taxon?	134 217 728	Not yet available	OBIS
29	Outliers:geography	Is the observation within six MADs from the distance to the centroid of this dataset?	268 435 456	Not yet available	OBIS

2. Assessment of the completeness and validity of information

Besides the basic information of a distribution record (what—where—by whom), the OBIS Schema can capture a lot of other species-related information. A number of the quality checks verify the completeness and soundness of different parts of information in a record. This includes traceability information—e.g. institution code and catalogue number—checking how detailed the date information is, verifying that a given date is possible and—if relevant—if the start date is always before the end date and the minimum depth is always smaller than or equal to the maximum depth.

A number of QC steps make it possible to distinguish between records that can be used as 'presence-only' or where actual counts are available. When a count is given, it is checked whether an indication of the sample size is documented, allowing users to recalculate the given values to a chosen unit. These QC flags give users the opportunity to e.g. only select those distribution records that have complete abundance information available or where the life stage is documented.

3. Taxonomic quality control

One of the most important quality checks within OBIS and EurOBIS is related to the given taxon names within a dataset. To quality check these names, (Eur)OBIS makes use of the World Register of Marine Species (WoRMS, WoRMS Editorial Board 2016, http://www.marinespecies.org) as the taxonomic standard. WoRMS is the most authoritative and comprehensive list of names of marine organisms, including information on synonymy. The host institute for WoRMS is the Flanders Marine Institute (VLIZ) in Belgium and the content of WoRMS is updated and validated by a world-wide network of taxonomic experts. Only by linking the given taxon names to a widely accepted marine taxonomic standard, such as WoRMS is it possible to rule out spelling variations and link synonyms to their currently accepted names within (Eur)OBIS. A thorough taxonomic standardization allows the grouping of distribution records in a reliable way for further analysis (Vandepitte et al., 2010).

4. Geographic quality control

As EurOBIS and OBIS are biogeographic information systems, verifying the geographic content is as important as verifying the taxonomic data. The geographic checks do not only include a 2D check—latitude and longitude—but they also evaluate the third dimension— depth—if documented in the dataset.

Several checks relate to the latitude–longitude fields within a given dataset (see Table 1). First of all, it is evaluated whether the coordinates are documented and if

the provided values are possible, i.e. be different from zero, be expressed as decimal values in the WGS84 format and fall within the valid boundaries (-90≤latitude≤+90 and -180≤longitude≤180). Although 0-0 is a marine position in the Gulf of Guinea (Atlantic Ocean), the odds of having sampled at that exact location is relatively small; All 0-0 cases in OBIS so far were referring to unknown positions, which have been auto-filled by zeros. As both data systems are marine, it is verified whether the sampling locations are located in the marine environment, being seas or oceans. Given the fact that they both receive coastal and estuarine datasets, a land mask accommodating for a 20 km buffer from the coastline (GSHHS, http://www.ngdc.noaa.gov/mgg/shorelines/gshhs.html) is taken into account, hence also including most of the estuarine areas. Although some datasets document the coordinate uncertainty or precision, this information has thus far not been taken into account in any of the quality control steps.

In nearly all cases, a dataset is accompanied by a detailed metadata description, including text information on the geographical range. Within the metadata information system used for EurOBIS, this geographical range information is coupled to Marine Regions (http://www.marineregions.org), a standard list of marine georeferenced place names and areas (Claus et al., 2014). Based on the available information and shape files within Marine Regions, a comparison is made between the location of the sampling points and the general geographical coverage mentioned in the metadata. If this does not correspond, the relevant sampling locations are flagged as possibly incorrect. When no metadata is available, this check cannot be performed and the record is evaluated as being correct. This check is not yet available on the OBIS database.

Within the marine environment, the relevance of information on sampling depth cannot be underestimated. Based on depth, it is possible to distinguish between e.g. planktonic and benthic observations or coastal and deepsea observations. Given its importance, it is valuable to evaluate if the given depth-value related to the species observation is a possible value. This assessment combines the given depth-values with their geographic coordinates and compares this to the General Bathymetric Chart of the Oceans (Anon., 2010). As not all depth values are registered with the same precision—and fluctuations exist due to e.g. tidal differences—a 100 m margin is taken into account when assigning a quality flag for this check. This margin should also largely account for the fact that the mean depth within a grid can potentially differ from the actual sampling depth, especially in topographically complex areas.

5. Outlier analysis

Next to the earlier documented QC steps that run both on EurOBIS and OBIS, global geographic and environmental outlier analyses were developed specifically for OBIS, generating 10 more QC flags. These additional outlier analyses use external environmental and geographical (depth) data to assess the credibility of a certain distribution record, when compared with the available distribution records within the checked dataset or within OBIS as a whole. Given the non-normal distribution of the environmental, depth and distance values of the sampling points, the following two robust outlier detection methods are used: (i) the absolute deviation from the median, with a limit at six times the median absolute deviation (MAD) (Davies & Gather, 1993; Leys et al., 2013) and (ii) an approach based on the Tukey box plot method, with boundaries at three times the interquartile range (IQR) (Acuna & Rodriguez, 2004). Although a value of three times MAD is already considered as conservative (Miller, 1991), setting the values for the rejection criteria is by definition a subjective decision (Leys et al., 2013). The values used for the QC flags are based on visual analysis of a subset of the OBIS database and on the fact that a point lying at 6xMAD or 3xIQR from the first or third quartile is considered an extreme outlier (Acuna & Rodriguez, 2004).

Six of the outlier checks are related to the environment: these checks compare the locality details of a record with depth, sea surface salinity (SSS) and sea surface temperature (SST) values extracted from the global grids of (1) GEBCO (www.gebco.net; The GEBCO_08 Grid, version 20100927), (2) ETOPO1 Global Relief Model (Amante & Eakins, 2009) and (3) MARSPEC (Ocean Climate Layers for Marine Spatial Ecology, Sbrocco & Barber, 2013), with the earlier explained decision criteria of 6xMAD and 3xIQR. The depth layers of these three global grids are combined and the average of the two most similar depth values is used to average out inconsistencies between the three bathymetric layers. It needs to be taken into account that due to the used resolution of these depth layers—30 arc-second for GEBCO_08 and MARSPEC and 1 arc-minute for ETOPO1 Global Relief Model—the calculated bathymetric values of the positions can significantly deviate from the values at the exact sampling position due to the resolution of the depth layers. These checks help identifying observations that (possibly) occur outside of their environmental range. The four geographic outlier procedures aim (i) to compare the orthodromic or great-circle distance between the actual sampling locations and the centroid of all sampling locations within a specific dataset and (ii) to compare the distance between the sampling location of a specific species record to the centroid of all the available sampling locations of that particular species within the OBIS

database. The quality flag is assigned taking into account the 3xIQR or 6xMAD boundaries. The centroid of a set of sampling points is defined as the point that minimizes the sum of squared geodesic distances between itself and each point in the set and it is calculated from all the initial records except those that have zero coordinates or coordinates that fall out of the valid boundaries for the coordinate reference system WGS84.

The outlier analyses aim to identify species documented outside of their expected ranges and to reveal possible errors in the taxonomic identification or the assigned latitude and longitude which were not identified through the record-level geographic QC steps, e.g. a missing minus sign to indicate South or West or accidental switching of latitude and longitude values.

Results

All distribution records within EurOBIS and OBIS have gone through the earlier described quality control steps. Within the OBIS database, at least 60% of the distribution records pass each individual QC step. For some QC steps, >90% of the records pass the enforced criteria (Fig. 1). A detailed look shows that the scores of the different OBIS nodes vary greatly (Fig. 2), indicating that the results of these QC procedures can provide valuable feedback to the data providers—to double check their data and possibly make corrections and additions—and users, to select the desired data from the system. For an overview of all datasets available within the OBIS database, we refer to http://www.vliz.be/en/imis?module=dataset&dasid=68.

The results show that 85% of the distribution records in OBIS can be used for species or genus specific analyses (Fig. 1). All nodes—and thus implicitly OBIS—seem to struggle with capturing the corresponding time zone of the given time at which the data were collected (QC13), which is valuable information when collating data from different time zones. Time and the corresponding time zone information is, e.g. highly relevant when comparing data from different regions and analysing the diurnal vertical migration patterns of, e.g. zooplankton species.

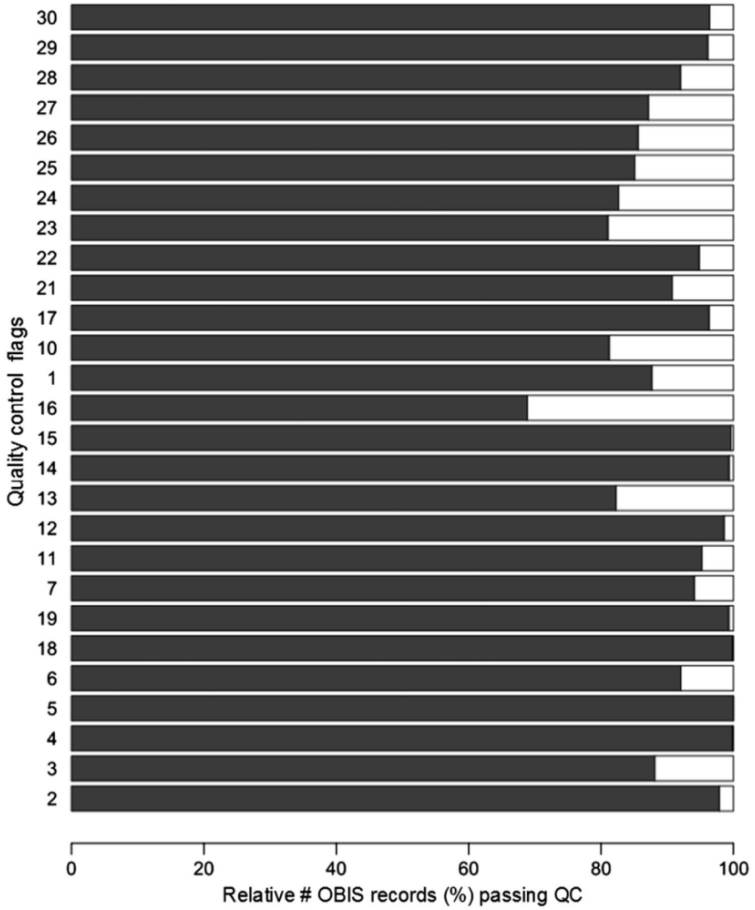


Figure 1. Relative number of records (%) that pass the individual QC steps within the OBIS database. The QC steps are listed in Table 1.

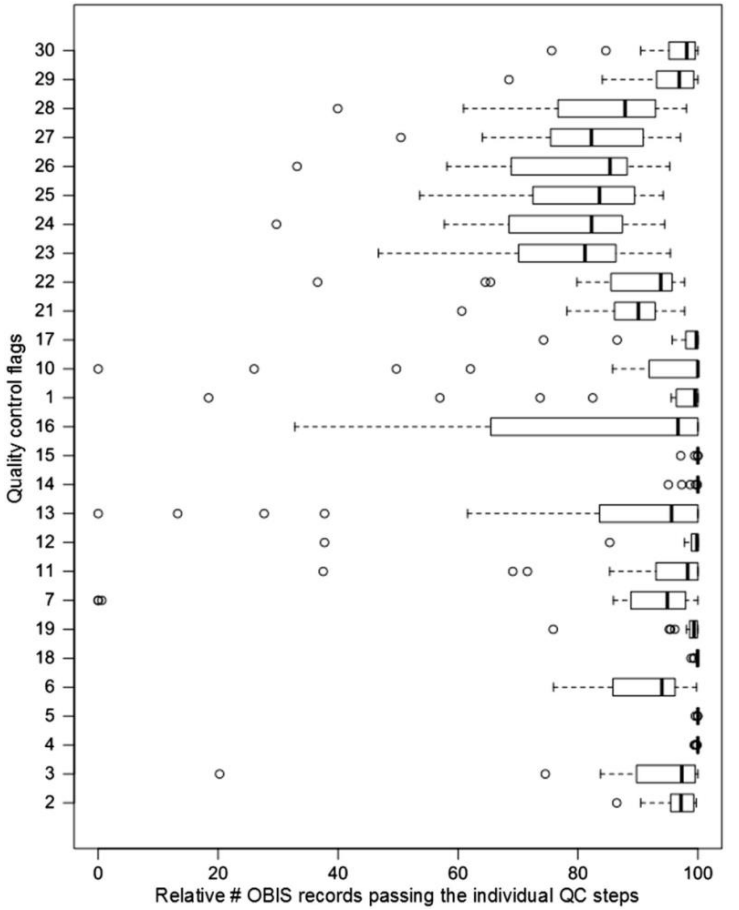


Figure 2. Box and whisker plot per QC step, showing the variability of quality and completeness (in percentage) of the distribution records within the 21 OBIS nodes.

When evaluating the records that contain actual counts (the number of observed individuals within each species) within the (Eur)OBIS database, it becomes clear that the most valuable piece of information—an indication of the sample size—is missing for a large number of records (QC16). As most counts are in essence meaningless without a sample size, this QC result shows that still a lot of work needs to be done to be able to use the count information.

Although the results of the individual QC steps can already give a lot of information on the possible usefulness of a record, it becomes even more useful when several QC steps are combined (Table 2). A selection of relevant QC steps can be made on database level, giving an indication of the distribution records within OBIS that comply to these criteria. In biodiversity research, scientists are specifically interested in geo-referenced species and/or genus data. When combining these selection criteria, almost 85% of the records would be fit for this purpose. The more stringent the criteria become, the fewer records will suit the postulated conditions. The number of suitable records diminishes significantly if one wants to make use of counts or abundance information instead of just presence information (QC16), indicating that this information is rather hard to capture and document within large integrated databases, such as e.g. OBIS.

Table 2. Overview of the number of records (absolute and relative) that pass specific combinations of QC steps, indicating their fitness for use in analysing research hypotheses. QC2: taxon name matched to the WoRMS; QC3: taxon level more detailed than family; QC4: coordinates different from zero; QC5: coordinates within possible boundaries; QC6: coordinates in sea or within 20 km coastline buffer; QC7: sampling year available and valid; QC16: count available, in combination with sample size information.

Combined QC steps	Positively evaluated OBIS records (#)	Positively evaluated OBIS records (%)
2-3-4-5	34 991 925	86.05
2-3-4-5-6	32 216 817	79.22
2-3-4-5-7	32 849 480	80.78
2-3-4-5-6-7	30 311 653	74.54
2-3-4-5-16	23 315 398	57.33
2-3-4-5-6-16	19 189 668	47.19
2-3-4-5-6-7-16	19 189 668	47.19

Two different approaches are used within the outlier analyses: the IQR and the MAD methodology. These two have been selected as they are widely used in outlier analyses. In general, the results of both QC procedures are similar. When they differ, the user can combine the results of these QC steps with other QC steps to come to a consensus approach on how to evaluate a specific record. Figures 3 and 4 illustrates

that the MAD and IQR approaches can differ, but that these differences are generally relatively small. If a record gets flagged as a possible outlier, some caution is still needed. Figure 3 represents the sampling locations of the dataset 'International Council for the Exploration of the Sea (ICES) Biological Community' (ICES, 2010), where the core of the locations is in the Baltic Sea and the other locations are indicated as geographic outliers. After consultation with the data management team at ICES, it became clear that the records in the Antarctic region were the result of a reporting problem in an old format, where positive latitudes were reported as negative. These errors are currently being fixed, and the correct data should soon be available. Possible issues with the Mediterranean, African mainland and Greenland records are not obvious and are still under investigation by ICES.

Fig. 4 shows all the distribution records of the Cirriped species Verruca stroemia available within OBIS and how they respond to the different geographic and environmental outlier analysis. The Supporting information gives an overview of the OBIS datasets containing Verruca stroemia distribution records. In the 'distance outlier analysis', all distribution records along the Norwegian coast, White Sea, Barents Sea and Mediterranean Sea are considered outliers, indicating the species would not occur there. Similar results come from the sea surface salinity (SSS) outlier analysis. Accepting these distribution records as true outliers should be backed up with expert knowledge, as these outliers might not be actual outliers, but e.g. the result of a skewed availability of data within the OBIS database or misidentifications in the field (see discussion).

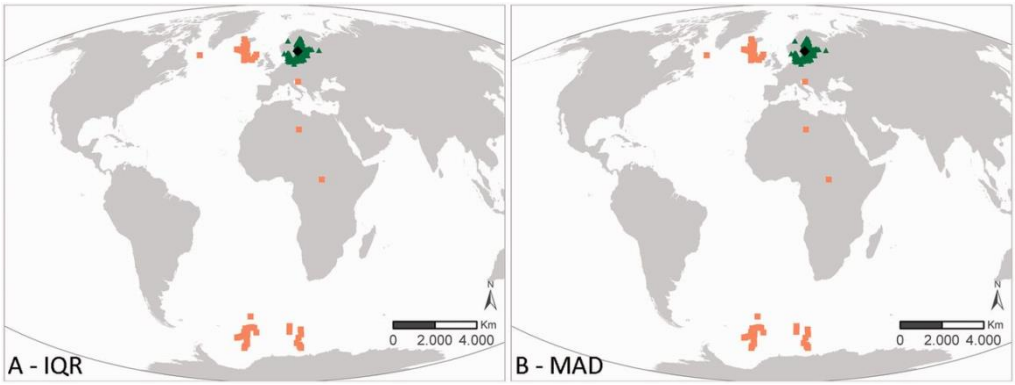


Figure 3. Results of the geographic outlier analysis on the dataset 'ICES Biological Community'. The left figure (A) represents the IQR approach, the right figure (B) represents the MAD approach. Black diamonds indicate the centroid of the investigated data, green triangles have been evaluated as OK, orange squares have been evaluated as possible outliers. In this case both the IQR and MAD identified the same points as outliers.

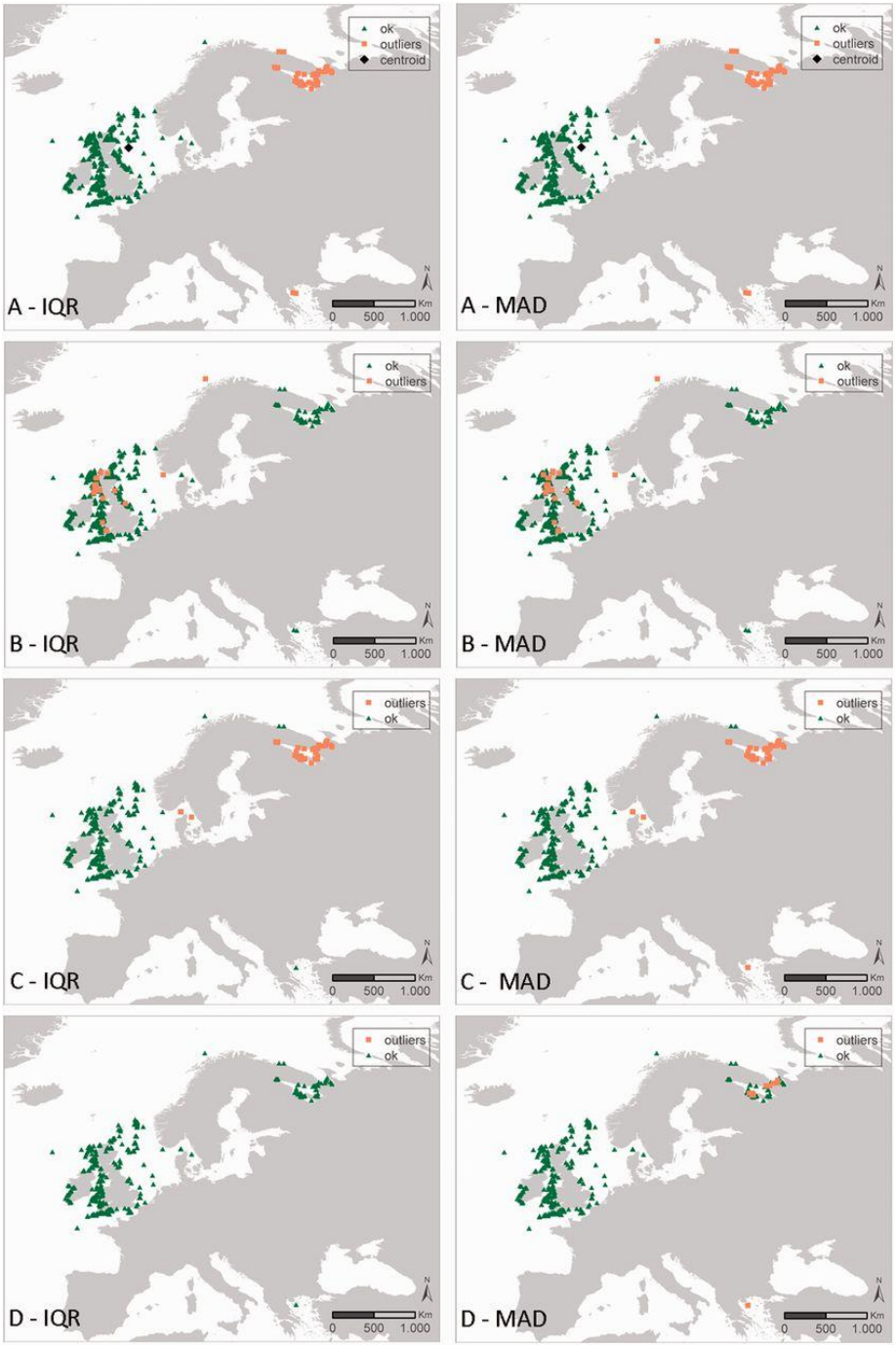


Figure 4. Results of the geographic and environmental outlier analysis of the species *Verruca stroemia* (Crustacea, Cirripedia). The left column represents the IQR approach, the right column represents the MAD approach. The different outlier analyses are A: geography, B: bathymetry, C: sea surface salinity (SSS) and D: sea surface temperature (SST). Black diamonds indicate the centroid of the investigated data (only for the geographic outlier analysis), green triangles have been evaluated as OK, orange squares have been evaluated as possible outliers.

Discussion

The quality flags assigned to each record provide an indication of the 'fitness for purpose' of a particular distribution record, helping both the user and the data provider in more objectively assessing the quality and completeness of a record and to draw conclusions from this. The majority of the quality flags do not have the intention to label a record as 'good' or 'bad', they just give an indication of the completeness and quality, helping the user in his or her decision to make use of a specific record or to reject it.

Users need to be aware of the fact that the results of the outlier analyses only provide an indication of the possible outlier character of a distribution record. Records flagged as an outlier are not necessarily true outliers: the distribution of a species can e.g. be unrelated to bathymetry, but highly dependent on temperature or salinity. A single outlier check might thus not clearly identify an outlier (Fig. 4), but combining the results of the different outlier checks can indicate with more certainty that a species observation is outside its suspected range (Fig. 5). In addition, knowledge on the actual environmental boundaries of species can help in identifying true outliers and filtering of the data. False positives in the species-based outlier detection can be the result of extremely uneven sampling such as for example data from museum collections. Some true positives on the other hand might not be actual outliers, but could be the first observations for a specific species in a geographical area where it was unknown to appear before. The latter could be the case in first observations of alien species that moved to a new area, and these records should be approached with caution. As the dataset-based outlier detection aims to flag possible errors in the geographic coordinates, this will only work well when the dataset is spatially restricted, e.g. if all samples have been taken in the same region such as the North Sea.

When wider geographical areas are covered within a dataset, this outlier detection is prone to giving false positives, e.g. due to a biased sampling effort in the available data. This is clearly the case for Verruca stroemia (Fig. 4): expert and literature consultation have confirmed that the Mediterranean outliers are true outliers, a consequence of misidentification (Young et al., 2003). In this case, the providers of these records will be contacted with the expert and literature information. The northern distribution records (Norwegian coast, White Sea, Barents Sea) are, however, validated by literature. In addition, the available depth values also confirmed the species occurs at a depth range from 0 to 548 m (43). Because different outlier analyses are available, it is recommended that users combine the results of these outlier QC checks with each other and with the results of the more

basic geography checks. All these combined will make the interpretation of the validity and fitness for use of the records.

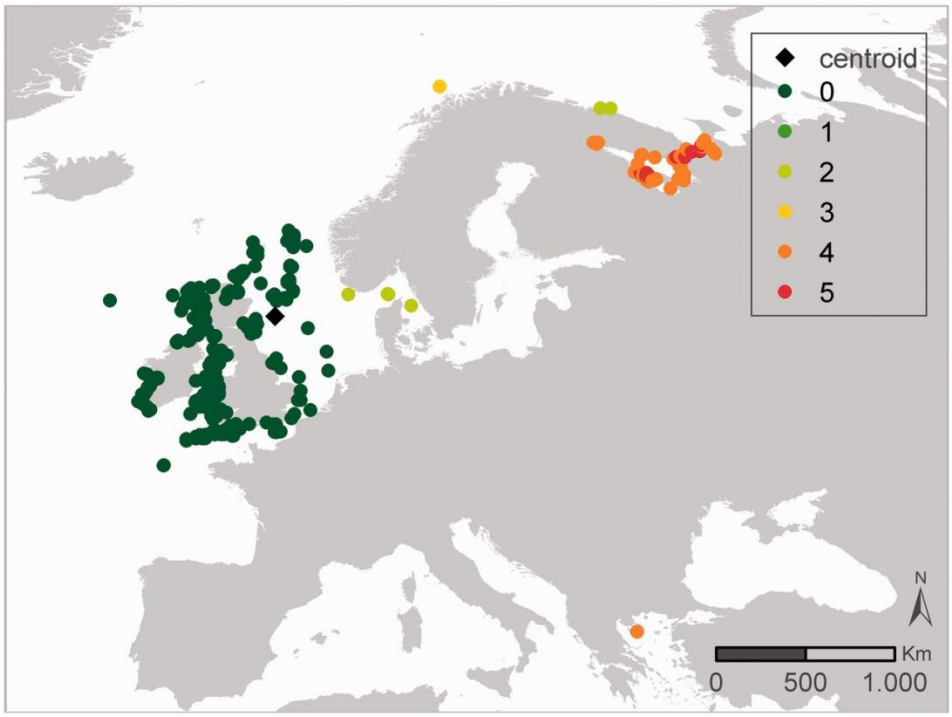


Figure 5. Synthesis map representing the combined results of the outlier analyses of Verruca stroemia from Figure 4. The scale represents the number of times a species distribution is seen as an outlier, when combining the eight outlier analyses—geography, bathymetry, Sea Surface Salinity (SSS) and Sea Surface Temperature (SST) SSS and SST according to the IQR and MAD approach—from Figure 4. The black diamond indicates the centroid of the investigated data.

Use-case 1: Quality controlled data available through EMODnet

As mentioned earlier, the results of the assigned quality control flags can be combined according to the required 'fitness for use' for the users, thereby generating the possibility to create specific filters on the available data within EurOBIS and OBIS. EMODnet Biology Portal (http://www.emodnet-biology.eu/) is already making use of such a filter, to offer a specific subset of EurOBIS data to its users. EurOBIS is the data engine behind the Biology Portal of EMODnet, meaning that the data part of the Biology Portal is driven by the EurOBIS data. It was, however, agreed that only those distribution data that comply with QC steps 2-3-4-5—related to taxonomy and basic geography—are offered to the users, thereby making a useful 'pre-selection' of the data. Through the portal, users can still see how many distribution records are available in the original dataset and how many have passed the postulated QC steps and are thus available. As of November 2014,

86% or 15.9 million of all the distribution records available in EurOBIS can be consulted through the EMODnet Biology Portal.

Use-case 2: Selection of QC steps available as web services through LifeWatch

As of 2012, EurOBIS is part of the central taxonomic backbone of LifeWatch, an E-Science European Infrastructure for Biodiversity and Ecosystem Research which aims at standardizing species data and integrating the distributed biodiversity repositories and operating facilities. Given the importance of standardization, interoperability and being able to assess the quality and completeness of the available data within LifeWatch, a number of the QC steps related to data format, taxonomy and geography that are currently running on the (Eur)OBIS database have been 'translated' to interactive, user-friendly web services (http://www.lifewatch.be/dataservices). By making use of these freely available data services, data providers, data managers and users are able to make a general assessment of the quality, completeness and fitness for use of their own biogeographic data by simply uploading them to the LifeWatch portal and selecting the QC steps they want to run on their data.

Future plans and possibilities

Currently, the QC steps are running automatically on both the EurOBIS and OBIS database. A selection of these QC steps is already available online through LifeWatch as a web service. The creation of a customized filter—a combination of several QC steps—is not yet available for the users. Customized filters on EurOBIS will become available through the EMODnet Data Portal, allowing users to define the necessary 'fitness for use' of the required data and to refine their search results accordingly. In the future, similar filter options will be developed on the OBIS data. The data download will then also include the corresponding QC flags. The results of the QC procedures currently stored in the database will be used to communicate with the data providers to improve both the quality and completeness of the available data. Specifically the outlier analyses will provide valuable information to improve the correctness of the data. Currently, newly added datasets are thoroughly analysed before they go online, and possible issues are communicated with the data provider immediately. On the other hand, a lot of data have been uploaded to the database before these QC procedures came into place. For these datasets, a communication plan will need to be worked out to discuss the quality control results with the providers, aiming for the highest possible return and improvement of the data quality and completeness. It is important to realize that for some—mostly

historical—datasets, the quality status will remain 'as is', e.g. when no additional information is available anymore and the original data provider is no longer around to deal with the identified issues.

Within WoRMS, the taxonomic information is currently being expanded with species attributes, such as whether a species belongs to the benthos or plankton, if a species is coastal or deep-sea, what the feeding method, average body size and life span is etc. Once these literature and expert-based traits have been sufficiently documented, they can be incorporated in the QC steps to offer an even higher quality standard to our users. For example, if WoRMS can distinguish between coastal and open ocean species, then this trait can be used as an additional check on the species distribution information: a coastal species (presumably) observed in the open ocean could then be flagged as a possibly incorrect record, drawing the attention of the users to this and letting them decide for themselves whether they want to include this record in their download or analysis or not.

Conclusion

The development and implementation of the described QC steps meets a need to be able to add quality flags to records and to filter out data based on user needs, taking into account the fitness for purpose of the available records. As an array of QC steps is available, users will be able to create specific filters on the data, answering to their specific data needs and requirements.

Although a number of the discussed QC steps are specifically designed to check data meant for EurOBIS and OBIS, a number of other checks can be used widely by the scientific community to quality control their own data before analysis, publication and data sharing. Offering these QC tools as online, user-friendly data services through LifeWatch (http://www.lifewatch.be) greatly enhances their overall usability for scientists worldwide and meets the needs of the (marine) scientific community to be able to standardize and quality check their data themselves.

Depending on user needs, more QC steps can be added in the future, or existing QC steps could be fine-tuned to better meet their requirements. The mining of a quality controlled, integrated database of different data sources can give insights in previously unexplored matters and offers the possibility to develop new or improved technologies related both to the quality of the data and the outcomes. It is, however, important to realize that the outlier QC results should be approached with due caution. Because the QC steps are automated, a critical analysis of these QC

results might be needed to draw the right conclusions on exclusion or inclusion of these records in certain analyses.

Acknowledgements

The European Ocean Biogeographic Information System (EurOBIS) is managed by the Flanders Marine Institute, with financial support from the Flemish Government. The development of the quality control procedures and making them available as online data services through the Belgian LifeWatch Portal is part of the Flanders Marine Institute (VLIZ) contribution to LifeWatch and is funded by the Hercules Foundation. Olivier De Clerck is indebted to EU FP7 ERANET (Project SEAS-EAR/INVASIVES).

The authors would like to thank Mike Flavell for his assistance in the implementation of the QC procedures on the OBIS database; Carlos Pinto of the ICES for giving feedback on the geographic outlier analysis of the ICES data; Robert van Syoc (California Academy of Sciences—CAS), WoRMS editor of the Cirripedia for his valuable input on the Verruca stroemia outlier analysis and the three anonymous referees for their constructive feedback on an earlier version of the manuscript.

Supporting information

Allen D., Beckett B., Brophy J., Costello M.J., Emblow C., Maciejewska B., McCrea M., Nash R., Penk M. & Tierney A. Marine species recorded in Ireland during field suveys by EcoServe, Ecological Consultancy Services Ltd. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=1947

Baranova, O.K, T.D. O'Brien, T.P. Boyer and I.V. Smolyar (2009). Plankton data. Chapter 16 in Boyer, T. P., J. I. Antonov , O. K. Baranova, H. E. Garcia, D. R. Johnson, R. A. Locarnini, A. V. Mishonov, T. D. O'Brien, D. Seidov, I. V. Smolyar, M. M. Zweng, 2009. World Ocean Database 2009. S. Levitus, Ed., NOAA Atlas NESDIS 66, U.S. Gov. Printing Office, Wash., D.C., 216 pp., DVDs. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=4099

CEFAS. - UK. Macrobenthos from English waters between 2000-2002. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=1681

Cochrane, S. (2001). Macrobenthos from the Norwegian waters. Akvaplan-niva, Norway. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=1856

Countryside Council for Wales. Marine Nature Conservation Review (MNCR) and associated benthic marine data held and managed by CCW. Countryside Council for Wales, Gwynedd, UK. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=657

Craeymeersh J., P. Kingston, E. Rachor, G. Duineveld, Carlo Heip, Edward Vanden Berghe, 1986: North Sea Benthos Survey. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=67

Dale Rostron. Marine records from Pembrokeshire Marine Species Atlas. Countryside Council for Wales, Gwynedd, UK. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=692

English Nature. Marine Nature Conservation Review (MNCR) and associated benthic marine data held and managed by English Nature. English Nature, Peterborough, UK. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=688

Fisheries Research Service, Marine Laboratory. Macrobenthos samples collected in the Scottish waters in 2001. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=1853

Flanders Marine Institute (VLIZ). Taxonomic Information System for the Belgian coastal area. 10 Aug 2004, Oostende, Belgium. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=82

Hellenic Centre For Marine Research, MedOBIS - Mediterranean Ocean Biogeographic Information System. Hellenic Centre for Marine Research; Institute of Marine Biology and Genetics; Biodiversity and Ecosystem Management Department, Heraklion, Greece. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=481

Mackie, A.S.Y., James, J.W.C., Rees, E.I.S., Darbyshire, T., Philpott, S.L., Mortimer, K., Jenkins, G.O. & Morando, A., 2006. The Outer Bristol Channel Marine Habitat Study. - Studies in Marine Biodiversity and Systematics from the National Museum of Wales. BIOMÔR Reports 4: 249 pp. & Appendix 228 pp. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=3068

Mackie, A.S.Y., P.G. Oliver, E.I.S. Rees, 1991: Biomôr 1 dataset. Benthic data from the Southern Irish Sea from 1989-1991. National Museum and galleries of Wales, Cardiff, UK. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=1600

Marine Conservation Society. Seasearch Marine Surveys. Marine Conservation Society, Ross-on-Wye, UK. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=746

Marine Ecological Surveys Ltd. - UK. Macrobenthos from the eastern English Channel in 1999 and 2001. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=1684

Naumov, A. Benthos of the White Sea. A database. White Sea Biological Station, Zoological Institute RAS. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=2769

Ostler, R. Marine Nature Conservation Review (MNCR) and associated benthic marine data held and managed by JNCC. Joint Nature Conservation Committee, Centre for Ecology and hydrology, Aberdeenshire, UK. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=621

Parr, J. Marine Life Information Network (MarLIN) marine survey data (Professional). Marlin, Collated Marine Life Survey Datasets, Marine Biological Association of the UK, Plymouth, UK. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=640

Picton, B.E., C.S. Emblow, C.C. Morrow, E.M. Sides, P. Tierney, D. McGrath, G. McGeough, M. McCrea,P. Dinneen, J. Falvey, S. Dempsey, J. Dowse, and M. J. Costello, 1999: Marine sites, habitats and species data collected during the BioMar survey of Ireland. Environmental Sciences Unit, Trinity College, Dublin, Ireland. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=345

Rees, H.L., Pendle, M.A., Waldock, R., Limpenny, D.S., Boyd, S.E. A comparison of benthic biodiversity in the North Sea, English Channel and Celtic Seas - Epifauna. Centre for Environment, Fisheries and Aquaculture Science; Burnham Laboratory, 12 Apr 2005, Essex, UK. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=505

Rees, H.L., Pendle, M.A., Waldock, R., Limpenny, D.S., Boyd, S.E. A comparison of benthic biodiversity in the North Sea, English Channel and Celtic Seas - Macroinfauna. Centre for Environment, Fisheries and Aquaculture Science; Burnham Laboratory, 12 Apr 2005, Essex, UK. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=3094

Rigby,P.R., B.Konar, T.Kato, K.Iken, H.Chenelot and Y.Shirayama (2005). NaGISA OBIS Dataset ver.1.. In: NaGISA . 2005. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=1983

Scottish Natural Heritage. Marine species data for Scottish waters held and managed by Scottish Natural Heritage, derived from benthic surveys 1993 to 2012. Scottish Natural Heritage, Edinburgh, UK. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=690

The Danish Biodiversity Information Facility, Marine Benthic Fauna List, Island of Læsø, Denmark. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=2038

The Norwegian Oil Industry Association, 2000: Offshore reference stations, Finnmark. The Norwegian Oil Industry Association (OLF), Akvaplan-niva and Det Norske Veritas, Norway. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=998

The Norwegian Oil Industry Association, 2002: Offshore reference stations, Norwegian/Barents Sea. The Norwegian Oil Industry Association (OLF), Akvaplan-niva and Det Norske Veritas, Norway. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=997

UK National Biodiversity Network, Countryside Council for Wales - Survey of North Wales and Pembrokeshire Tide Influenced Communities. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=1883

UK National Biodiversity Network, Marine Biological Association - DASSH Data Archive Centre Academic surveys. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=1890

UK National Biodiversity Network, Marine Biological Association - DASSH Data Archive Centre expert sighting records. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=1885

UK National Biodiversity Network, Marine Biological Association - DASSH Data Archive Centre volunteer sightings records. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=1891

Wilkinson, S. Marine benthic dataset (version 1) commissioned by UKOOA. Joint nature Conservation Committee, Peterborough, UK. Metadata: http://www.vliz.be/en/imis?module=dataset&dasid=645

Chapter 3

sdmpredictors: an R package for species distribution modelling predictor datasets

Samuel Bosch[1,2], Lennert Tyberghein[1], Olivier De Clerck[2]

[1]*Flanders Marine Institute (VLIZ), InnovOcean site, Wandelaarskaai 7, 8400 Ostend, Belgium*
[2]*Research Group Phycology, Biology Department, Ghent University, Krijgslaan 281/S8, 9000 Ghent, Belgium*

Unpublished manuscript.

Abstract

sdmpredictors is an open source R package which allows the end user to download terrestrial and marine environmental layers for the past, current and future climates. *sdmpredictors* contains metadata, statistics and pairwise correlations for the available datasets and layers. These correlations between predictors can be subsequently grouped and plotted. Currently *sdmpredictors* contains geophysical, biotic and climate data from WorldClim, ENVIREM, Bio-ORACLE and MARSPEC at 5 arcmin resolution and in the Behrmann equal area projection with a resolution of 7 kilometres.

Introduction

Species distribution modelling (SDM) is a commonly used tool in ecology and conservation biology. Correlative species distribution models relate species occurrences and (pseudo-)absence data to environmental predictor variables, based on statistically derived response surfaces (Guisan & Thuiller, 2005). Coinciding with the persistent interest in SDM, in the last 20 years numerous R packages related to SDM have been released. These include packages for downloading, checking and thinning occurrences (Aiello-Lammens et al., 2015; Chamberlain et al., 2016a; Provoost et al., 2016; Robertson et al., 2016), downloading the WorldClim environmental dataset (Hijmans et al., 2016; August et al., 2017), fitting models with various algorithms (Liaw & Wiener, 2002; Royle et al., 2012; Golding & Purse, 2016) and packages providing a fully integrated framework for SDM (Thuiller et al., 2009; Hijmans et al., 2016; Naimi & Araújo, 2016; August et al., 2017). With *sdmpredictors* we aim to complement these R packages by providing an easy to use interface for the acquisition of uniform and compatible terrestrial and marine predictors from different datasets for the past, current and future climate layers. It allows the end user to easily discover and use the different available layers from different predictor datasets.

Package description

sdmpredictors allows you to query the metadata for datasets ('list_datasets') and the environmental layers ('list_layers'). After selecting the required current climate layers they can be downloaded and loaded into the R session using the 'load_layers' function by providing the layer codes. Once layers are loaded into R they can be passed to all functions expecting a RasterStack with environmental data such as 'extract' from *raster* (Hijmans, 2016), 'BIOMOD_FormatingData' from *biomod2 (Thuiller et al., 2009)*, *LocalRaster* module in ZOON (August et al., 2017) and many more.

In order to load paleoclimatic and future climate layers a set of functions links current climate layers to past and future climate layers ('get_paleo_layers' or 'get_future_layers') or list out the available paleoclimatic and future climate layers ('list_layers_paleo' or 'list_layers_future'). After which the same 'load_layers' function can be used to actually download the data.

With the 'layer_stats' function various summarizing layer statistics like minimum, first quantile, median, third quantile, maximum, median absolute deviation, mean and standard deviation can be queried. The 'layers_correlation' function allows one

to query the Pearson correlation coefficient between two or more layers. Following the suggestion of Dormann et al. (2013), to avoid including heavily correlated predictors in one SDM, we provide the 'correlation_groups' function, which groups predictors based on their correlation. Correlations for cropped versions of the predictors or between externally sourced predictors can be calculated with the 'pearson_correlation_matrix' function.

Finally, citations for the used datasets and layers can be obtained with the 'dataset_citations' and 'layer_citations' functions, respectively.

Integrated datasets

Currently data layers are available both as Behrmann equal area projected rasters with a 7 km resolution and as 5 arcminutes unprojected rasters. For the terrestrial environment we added the WorldClim (Hijmans et al., 2005) and ENVIREM (Title & Bemmels, 2017) datasets and for the marine environment we included Bio-ORACLE (Tyberghein et al., 2012) and MARSPEC (Sbrocco & Barber, 2013). An overview of these datasets can be found in Table 1. The included datasets all represent multiyear aggregated data from interpolated *in situ* data and satellite observations of the Earth's surface. For all of datasets past and future climate data were added when available. This is by no means a fixed list and we encourage end users to suggest new datasets for inclusion in *sdmpredictors*.

Usage

In Supporting information we provide an example use case where *sdmpredictors* is used to provide environmental data for modelling the distribution of *Dictyota diemensis* Sonder ex Kützing, one of the species from the MarineSPEED benchmark dataset (Chapter 4). Additionally, the data provided by *sdmpredictors* can also be used for numerous other applications, including the generation of virtual species (Duan et al., 2015; Leroy et al., 2016), measuring niche overlap (Broennimann et al., 2012), linking the environment with species richness and biogeographic structure (Tittensor et al., 2010; Belanger et al., 2012) and modelling species abundance and population dynamics (Pearce & Boyce, 2006; Pratheepa et al., 2016). In summary, this package provides users with a set of functions for obtaining and using environmental predictor datasets for the past, current and future climate within R.

Table 1. Overview of the datasets included in *sdmpredictors*. For an up to date list use the function 'list_datasets'.

Dataset	Description
WorldClim	WorldClim is a set of global terrestrial climate layers. It has average monthly climate data for minimum, mean, and maximum temperature and for precipitation for 1960-1990. Additionally it contains a set of bioclimatic variables that are derived from the monthly temperature and rainfall values. They represent annual trends, seasonality and extreme or limiting environmental factors.
ENVIREM	The ENVIREM dataset is a set of 16 climatic and 2 topographic variables that can be used in modelling species' distributions. The strengths of this dataset include their close ties to ecological processes, and their availability at a global scale, at several spatial resolutions, and for several time periods. The underlying temperature and precipitation data that went into their construction comes from the WorldClim dataset (www.worldclim.org), and the solar radiation data comes from the Consortium for Spatial Information (www.cgiar-csi.org). The data are compatible with and expand the set of variables from WorldClim v1.4 (www.worldclim.org).
Bio-ORACLE	Bio-ORACLE is a set of GIS rasters providing marine environmental information for global-scale applications. It offers an array of geophysical, biotic and climate surface data derived from satellite data or interpolated from in situ data.
MARSPEC	MARSPEC is a set of high resolution climatic and geophysical GIS data layers for the world ocean. Seven geophysical variables were derived from the SRTM30_PLUS high resolution bathymetry dataset. These layers characterize the horizontal orientation (aspect), slope, and curvature of the seafloor and the distance from shore. Ten "bioclimatic" variables were derived from NOAA's World Ocean Atlas and NASA's MODIS satellite imagery and characterize the inter-annual means, extremes, and variances in sea surface temperature and salinity.

To cite sdmpredictors or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for 'version 0':

Bosch S., Tyberghein, L. and De Clerck, O. 2017. sdmpredictors: an R package for species distribution modelling predictor datasets. Version 0.

Acknowledgements

The research was carried out with financial support from the ERANET INVASIVES project (EU FP7 SEAS-ERA/INVASIVES SD/ER/010) and financial, data & infrastructure support provided by VLIZ as part of the Flemish contribution to the LifeWatch ESFRI.

Data accessibility

The *sdmpredictors* R package is available on CRAN and at https://github.com/lifewatch/sdmpredictors.

Supporting information

Here we detail a sample application where *sdmpredictors* is used to provide environmental data for modelling the distribution of *Dictyota diemensis* Sonder ex Kützing, one of the species from the MarineSPEED benchmark dataset (Chapter 4). The distribution of *D. diemensis* is restricted to Australia and New Zealand, but as only Australian distribution records are available we restricted ourselves for this use case to the Australian range (Womersley, 1987; Adams, 1994). We first start with exploring the available datasets and layers. Followed by the download of a set of 5 marine layers (salinity, sea surface temperature mean and range, bathymetry and shore distance) from Bio-ORACLE and MARSPEC. These are subsequently clipped with the shape of the Australian Exclusive Economic Zone using the *raster* and *mregions* packages (Chamberlain et al., 2016b; Hijmans, 2016). Secondly statistics and correlations for both the global and Australian data are inspected and visualized. For the correlation plot we additionally used the *ggplot2* and *cowplot* packages (Wickham et al., 2016; Wilke & Wickham, 2016). Since no predictors are grouped in a correlation group (Pearson correlation > 0.7) we used all predictors for building the SDM. We downloaded occurrences using *marinespeed* (Bosch et al., 2017), which are then used to create an SDM using ZOON (August et al., 2017). Finally the citations for the used layers are printed. For this application we used the Behrmann equal-area projected layers which required the projection of extents and occurrence points, avoiding oversampling of higher latitudes (Elith et al., 2011).

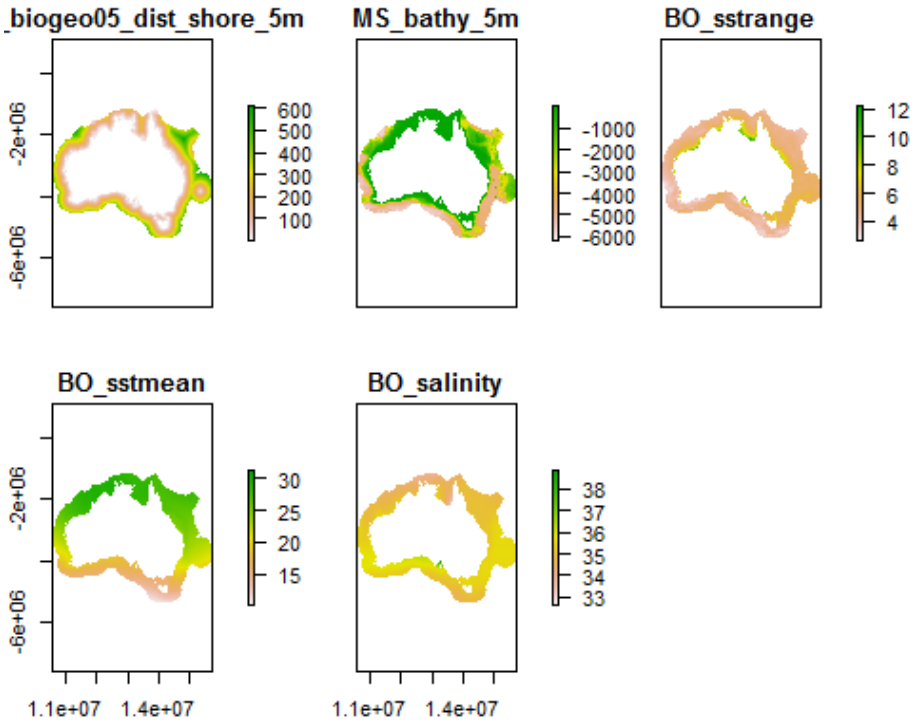
```
library(sdmpredictors)
library(mregions)
library(zoon)
# Inspect the available marine datasets and layers
datasets <- list_datasets(terrestrial = FALSE, marine = TRUE)
View(datasets[,c("dataset_code", "description")])
```

dataset_code	description
Bio-ORACLE	Bio-ORACLE is a set of GIS rasters providing marine environmental information for global-scale applications. It offers an array of geophysical, biotic and climate data at a spatial resolution 5 arcmin (9.2 km) in the ESRI ascii format.
MARSPEC	MARSPEC is a set of high resolution climatic and geophysical GIS data layers for the world ocean. Seven geophysical variables were derived from the SRTM30_PLUS high resolution bathymetry dataset. These layers characterize the horizontal orientation (aspect), slope, and curvature of the seafloor and the distance from shore. Ten "bioclimatic" variables were derived from NOAA's World Ocean Atlas and NASA's MODIS satellite imagery and characterize the inter-annual means, extremes, and variances in sea surface temperature and salinity. These variables will be useful to those interested in the spatial ecology of marine shallow-water and surface-associated pelagic organisms across the globe. Note that, in contrary to the original MARSPEC, all layers have unscaled values.

```
layers <- list_layers(datasets)
View(layers[1:2,c("dataset_code", "name", "description",
                  "primary_type")])
```

dataset_code	name	description	primary_type
Bio-ORACLE	Calcite (mean)	Calcite concentration indicates the mean concentration of calcite (CaCO3) in oceans.	Satellite (Aqua-MODIS), seasonal climatologies
Bio-ORACLE	Chlorophyll A (maximum)	Chlorophyll A concentration indicates the concentration of photosynthetic pigment chlorophyll A (the most common "green" chlorophyll) in oceans. Please note that in shallow water these values may reflect any kind of autotrophic biomass.	Satellite (Aqua-MODIS), monthly climatologies

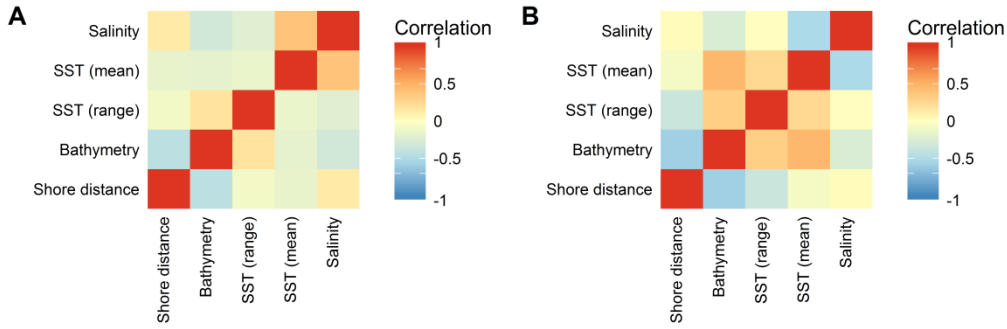
```
# Load equal area rasters, crop with the shape of the Australia EEZ
layercodes <- c("MS_biogeo05_dist_shore_5m", "MS_bathy_5m",
                "BO_sstrange", "BO_sstmean", "BO_salinity")
env <- load_layers(layercodes, equalarea = TRUE)
eez <- mregions::mr_shp("MarineRegions:eez", maxFeatures = NULL,
              filter = "Australian Exclusive Economic Zone")
eez <- sp::spTransform(eez, equalareaproj)
australia <- raster::crop(env, extent(eez))
australia <- raster::mask(australia, eez)
plot(australia)
```



```
# Compare statistics between original and Australian bathymetry
rbind(layer_stats("MS_bathy_5m"),
      calculate_statistics("Bathymetry Australia",
                           raster(australia, layer = 2)))
```

	layer_code	minimum	q1	median	q3	maximum	mad	mean	sd	moran	geary
171	MS_bathy_5m	-10493	-4865	-4082	-2984	-1	1313.584	-3661.049	1644.869	0.9728919	0.0096978
0%	Bathymetry Australia	-6163	-4377	-1868	-85	-1	2682.023	-2222.555	1987.391	0.9736722	0.0053917

```
# Compare correlations between predictors, globally and for Australia
prettynames <- list(BO_salinity="Salinity",
                    BO_sstmean="SST (mean)",
                    BO_sstrange="SST (range)",
                    MS_bathy_5m="Bathymetry",
                    MS_biogeo05_dist_shore_5m = "Shore distance")
p1 <- plot_correlation(layers_correlation(layercodes), prettynames)
australian_correlations <- pearson_correlation_matrix(australia)
p2 <- plot_correlation(australian_correlations, prettynames)
cowplot::plot_grid(p1, p2, labels=c("A", "B"), ncol = 2, nrow = 1)
```



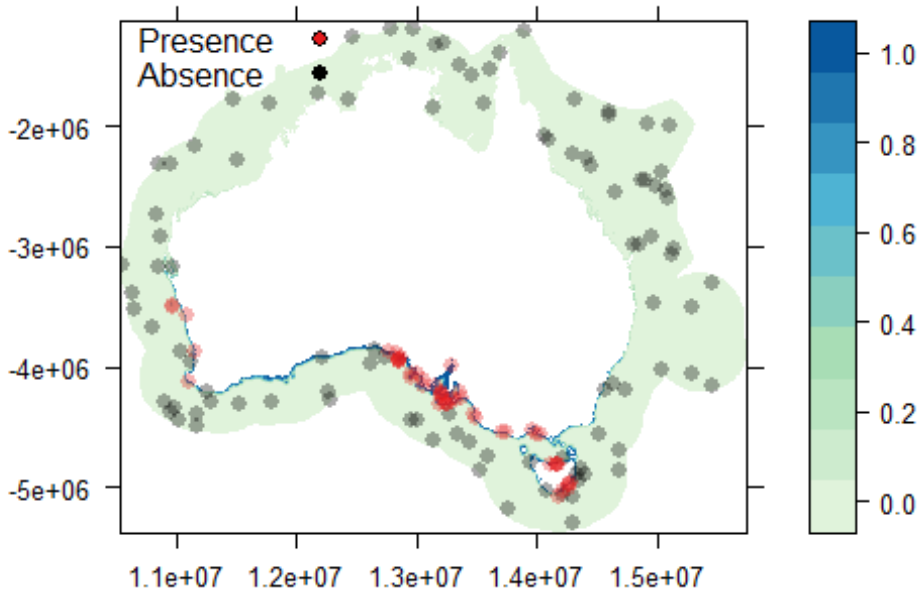

```
correlation_groups(australian_correlations, max_correlation = 0.7)

## [[1]]
##   MS_biogeo05_dist_shore_5m
## "MS_biogeo05_dist_shore_5m"
##
## [[2]]
##   MS_bathy_5m
## "MS_bathy_5m"
##
## [[3]]
```

```
##   BO_sstrange
## "BO_sstrange"
##
## [[4]]
##   BO_sstmean
## "BO_sstmean"
##
## [[5]]
##   BO_salinity
## "BO_salinity"
```

```
# Fetch occurrences and prepare for ZOON
occ <- marinespeed::get_occurrences("Dictyota diemensis")
points <- SpatialPoints(occ[,c("longitude", "latitude")],
                        lonlatproj)
points <- spTransform(points, equalareaproj)
occfile <- tempfile(fileext = ".csv")
write.csv(cbind(coordinates(points), value=1), occfile)
# Create SDM with ZOON
wf <- workflow(
  occurrence = LocalOccurrenceData(
    occfile, occurrenceType="presence",
    columns = c("longitude", "latitude", "value")),
  covariate = LocalRaster(stack(australia)),
  process = OneHundredBackground(seed = 42),
  model = LogisticRegression,
  output = PrintMap)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```



```r
# Layer citations
citations <- layer_citations(layercodes, astext=FALSE)
for(citation in citations) {
  print(citation, style="Bibtex")
}
```

```
## @Article{Bio-ORACLE,
##   author = {Lennert Tyberghein and Verbruggen Heroen and Klaas Pauly and
Charles Troupin and Frederic Mineur and Olivier {De Clerck}},
##   title = {Bio-ORACLE: a global environmental dataset for marine species
distribution modelling},
##   journal = {Global Ecology and Biogeography},
##   year = {2012},
##   volume = {21},
##   number = {2},
##   pages = {272-281},
##   doi = {10.1111/j.1466-8238.2011.00656.x},
## }
## @Article{MARSPEC,
##   author = {Elizabeth J. Sbrocco and Paul H. Barber},
##   title = {MARSPEC: ocean climate layers for marine spatial ecology},
##   year = {2013},
##   volume = {94},
##   number = {4},
##   pages = {979},
##   journal = {Ecology},
##   doi = {10.1890/12-1358.1},
## }
```

Chapter 4

In search of relevant predictors for marine species distribution modelling using the MarineSPEED benchmark dataset

Samuel Bosch[1,2], Lennert Tyberghein[1], Klaas Deneudt[1], Francisco Hernandez[1] and Olivier De Clerck[2]

[1]*Flanders Marine Institute (VLIZ), InnovOcean site, Wandelaarskaai 7, 8400 Ostend, Belgium*
[2]*Research Group Phycology, Biology Department, Ghent University, Krijgslaan 281/S8, 9000 Ghent, Belgium*

Submitted in March 2017.

Abstract

Aim

Ideally, datasets for species distribution modelling (SDM) contain evenly sampled records covering the entire distribution of the species, confirmed absences and auxiliary ecophysiological data allowing informed decisions on relevant predictors. Unfortunately, these criteria are rarely met for marine organisms for which distributions are too often only scantly characterized and absences generally not recorded. Here, we investigate predictor relevance as a function of modelling algorithms and settings for a global dataset of marine species. Furthermore, we promote the usage of a standardized benchmark dataset (MarineSPEED) for methodological SDM studies.

Location

Global marine.

Methods

We selected well studied and identifiable species from all major marine taxonomic groups. Distribution records were compiled from public sources (e.g. OBIS, GBIF, Reef Life Survey) and linked to environmental data from Bio-ORACLE and MARSPEC. Using this dataset, predictor relevance was analysed under different variations of modelling algorithms, numbers of predictor variables, cross-validation strategies, sampling bias mitigation methods, evaluation methods and ranking methods. SDMs for all combinations of predictors from 8 correlation groups were fitted and ranked, from which the top five predictors were selected as the most relevant.

Results

We collected two million distribution records from 514 species across 18 phyla and made them available with associated environmental data and cross-validation splits through the R package *marinespeed* and at http://marinespeed.org. Mean sea surface temperature and calcite are respectively the most relevant and irrelevant predictors. A less clear pattern was derived from the other predictors. The biggest differences in predictor relevance were induced by varying the number of predictors, the modelling algorithm and the sample selection bias correction.

Main conclusions

While temperature is a relevant predictor of global marine species distributions, considerable variation in predictor relevance is linked to the SDM setup. Future methodological SDM studies should consider the use of a benchmark dataset.

Introduction

Species distributions are increasingly modelled for conservation and ecological purposes. A better understanding of the mechanisms shaping species distributions allows for more accurate predictions of the future distribution of species in a rapidly changing world (Franklin, 2009). Climatological conditions are currently changing at an unprecedented rate and anthropogenic activities displace species out of their native area across the globe resulting in biological invasions (Walther et al., 2009).

A mechanistic link between the abiotic factors and the species distributions is traditionally gleaned from physiological studies subjecting individuals to various environmental conditions and assessing their reaction norms. However, not all species lend themselves equally well to ex situ experiments. Also, the experimental setup may only approximate realistic environmental conditions to a limited degree. Furthermore, such physiological studies typically require prior knowledge on the ecological factors governing the distribution ranges (Kearney & Porter, 2009). Given these difficulties, species distribution modelling (SDM), alternatively known as Ecological Niche Modelling (ENM), offers an attractive alternative (Elith et al., 2010). SDM correlates species occurrences, and optionally absences, with environmental data to create an estimation of the ecological niche and a projection in geographic space of this niche (Austin, 2002). The obvious advantage of correlative SDMs is that they require little knowledge of the mechanistic links between organisms and their environments. Thanks to the availability of an increasing number of online distribution records (e.g. OBIS, GBIF), pre-processed environmental data layers (e.g. Worldclim, Climond, Bio-ORACLE, MARSPEC) and modelling algorithms accessible through various statistical packages, SDM has become a widely applied technique in ecology and conservation biology (Pacifici et al., 2017).

Despite this, studies on general SDM theory and methodology mostly focus on the terrestrial environment (reviewed in Franklin 2009; Elith & Leathwick 2009; Peterson et al. 2011). A minority of papers specifically address distribution modelling methods in the marine environment: presence-only algorithms (Cheung et al., 2008; Ready et al., 2010; Beaugrand et al., 2011), algorithm comparisons (MacLeod et al., 2008; Palialexis et al., 2011; Šiaulys & Bučas, 2012), 3D modelling (Bentlage et al., 2013), rare species (Stirling et al., 2016), joint SDMs (Torres et al., 2008), ensemble modelling (Downie et al., 2013), scale effects (Pittman & Brown, 2011; Nyström Sandman et al., 2013), null models (Merckx et al., 2011), model selection (Verbruggen et al., 2013), pseudo-absence generation (Huang et al., 2011; Coro et al., 2016) and predictor datasets (Tyberghein et al., 2012; Sbrocco & Barber, 2013).

Although the importance of selecting biologically relevant predictors, and its impact on model uncertainty and transferability has been highlighted by several studies (Araújo & Guisan, 2006; Barry & Elith, 2006; Synes & Osborne, 2011; Braunisch et al., 2013; Verbruggen et al., 2013; Petitpierre et al., 2017) to date no comprehensive study on the relevance of the predictors of marine species distributions across taxa has been performed. But, note that Bradie & Leung (2016), in their meta-analysis on variable importance from MaxEnt SDMs, included a limited set of marine species. Bradie & Leung (2016) found that temperature and to a smaller extent bathymetry and salinity contributed the most to marine species distribution models. While the impact of geographic scale, algorithm and pseudo-absence selection on the importance of predictors have been addressed to some degree (VanDerWal et al., 2009; Elith et al., 2010; Nyström Sandman et al., 2013; Bucklin et al., 2015) the impact of these and other aspects of SDM have not been studied on a global scale.

In this study, we created the Marine SPEcies with Environmental Data (MarineSPEED) dataset. This benchmark dataset, containing distribution records belonging to 514 well-studied taxa with a broad taxonomic, climatologic and geographic diversity, is used to investigate marine predictor relevance under an array of modelling parameters and algorithms. With this, we aim to answer two questions: (1) what are the most relevant predictors of marine species distributions and (2) which part of the SDM process impacts the relevance of predictors the most. Additionally, this study aims to promote the usage of benchmark datasets in methodological SDM studies as this allows for reproducible and comparable results.

Methods

Species data

For the marine species benchmark dataset we selected species from an array of taxonomic groups, climatological preferences and distribution patterns. We aimed to include species that are well-studied in terms of their distribution and that often would classify as iconic species. For a species to be considered we required the availability of at least 100 distribution records in public databases.

Species distribution records were collected from the Ocean Biogeographic Information System (OBIS; iobis.org, accessed February 2016), from the Global Biodiversity Information Facility (GBIF; gbif.org, accessed January 2016), the Reef Life Survey (RLS; reeflifesurvey.com, accessed February 2016) and for a few species via personal communications. For downloading the records from OBIS and GBIF the R (R Core Team, 2016) clients *robis* (Provoost et al., 2016) and *rgbif* (Chamberlain et

al., 2016a) were used, respectively. A list of data sources is found in Appendix S1 in Supporting information. The distribution records were subsequently filtered until only one record remained in each cell of an equal-area grid with a per cell area of 25 square kilometres. This step eliminates duplicated records from different data sources and limits the number of records from repeated sampling events in the same area. We also removed records located within the land mask of the environmental data. Finally the distributions for all species were visually inspected and cross-checked with available distribution information in order to eliminate erroneous records. The amount of sample selection bias was assessed by visually comparing the spread of the occurrence records with the distribution range of the species and attributing a score ranging from 1 (low bias) to 5 (high bias).

We collected for each species taxonomic and functional group information from the World Register of Marine Species (WoRMS Editorial Board, 2016). The 'functional group' trait divides species into three groups reflecting their habitat: benthos, nekton and plankton (zooplankton and phytoplankton). For species lacking trait data in WoRMS, this information was derived from FishBase (Froese et al., 2017) and SeaLifeBase (Palomares et al., 2017) whereby all seafloor associated species were classified as benthos (i.e. sessile, reef-associated or demersal species), other free swimming species as nekton and drifting species as plankton. Additionally, we identified the latitudinal zones ('polar', 'temperate', 'tropical') for each distribution range. To do this, we checked for the presence of at least five per cent of all occurrence records of a species in each latitudinal zone of the marine ecoregions classification by Spalding et al. (2007). Lastly, species were categorized as oceanic if more than five per cent of their records are located outside the marine ecoregions. Else, species were considered as neritic.

Environmental data

The distribution records in the MarineSPEED dataset were linked to all 71 monthly and annual environmental variables for the current climate available from Bio-ORACLE (Tyberghein et al., 2012) and MARSPEC (Sbrocco & Barber, 2013) with a spatial resolution of 5 arcmin using the R package *sdmpredictors* (Bosch et al., 2016). This environmental data includes variations of sea surface temperature, salinity, bathymetry, nutrients and other predictors of marine species distributions.

Background data

Most presence-only SDM methods use background or pseudo-absence points for building models (Franklin, 2009). In order to facilitate the reproducibility of different studies using MarineSPEED we included a set of 20.000 randomly sampled background points in the benchmark dataset. We also created a second set of target-group background points by randomly sampling 20.000 points from the full

set of distribution records. The latter show the same bias as the occurrence records and therefore can be used to mitigate the effect of sample selection bias on presence-only species distribution models (Phillips et al., 2009; Kramer-Schadt et al., 2013; Syfert et al., 2013).

Cross-validation splits

Cross-validation (CV) is a widespread strategy used to perform model selection while avoiding under- and overfitting models (Arlot & Celisse, 2010). We prepared CV folds for the species and background data using three different strategies. As a first strategy we partitioned the data randomly in five folds (random CV). This strategy is easy to perform but has as disadvantage that it results in an overestimated performance of the model because training and validation points selected from nearby locations will be dependent due to the effect of spatial autocorrelation (Bahn & McGill, 2007; Hijmans, 2012; Roberts et al., 2016). As CV only avoids overfitting when training samples are independent from the validation samples this generally leads to the selection of complex models with poor transferability (Arlot & Celisse, 2010; Verbruggen et al., 2013; Petitpierre et al., 2017). The second (disc-based CV) and third (grid-based CV) splitting strategies take into account the spatial nature of the data. The 5-fold disc-based strategy randomly samples a starting point and subsequently selects the nearest one fifth of all distribution records to get the first fold. Then the distribution record furthest away from the starting point is used as a new starting point and the nearest one fifth of the distribution records are included to create the second fold. This process is repeated five times until all records are assigned to a fold. For the 4-fold grid-based strategy records are split into two sets based on their longitude using a random meridian as a dividing line. Then these two halves are separately split in two equal parts using parallels. Additionally, 9-fold grid-based sets were created by using two meridians and parallels for splitting instead of one. By combining the disc- or grid-based CV strategies with the pairwise distance sampling method proposed by Hijmans (2012) to select the pseudo-absence points for the test set spatial sorting bias was eliminated and thus the effect of spatial autocorrelation on the performance evaluation supressed (Bahn & McGill, 2007; Roberts et al., 2016). In order to remove false negatives in the training sets of the spatial cross-validation sets we excluded background points from the training sets that are within 200 km of test occurrences.

Predictor relevance

In order to find out which predictors are most relevant for the set of species in MarineSPEED we ranked distribution models fitted for all combinations of predictors from multiple correlation groups. In addition, we added variation at the different steps of the model creation to assess the variability in predictor relevance under different model setups (Fig. 1).

Following the methodology from Barbet-Massin & Jetz (2014), who identified relevant predictors of bird distributions, distributions were modelled for all combinations of three, four and seven environmental predictors selected from eight correlation groups. After filtering the initial set of 68 predictors down to 19 predictors based on a Pearson product moment correlation coefficient larger than 0.95 we created correlation groups with the R package *sdmpredictors* by grouping all predictors for which some or all of the predictors have an absolute Pearson product moment correlation coefficient larger than 0.7 (Dormann et al., 2013; Barbet-Massin & Jetz, 2014). This resulted in 8 correlation groups of which 6 predictors form a group on their own (shore distance, bathymetry, SST (range), calcite, salinity, pH), 7 predictors belong to the "Chlorophyll a group", grouping chlorophyll a and diffuse attenuation (mean, minimum, maximum and/or range) related variables. The last 6 predictors form the "SST group" with variations of sea surface temperature (SST), photosynthetically active radiation (PAR), phosphate, nitrate and silicate. For a full overview of the different environmental predictors used and the correlation group they belong to we refer to Fig. 2 and to Table S1 in Appendix S3.

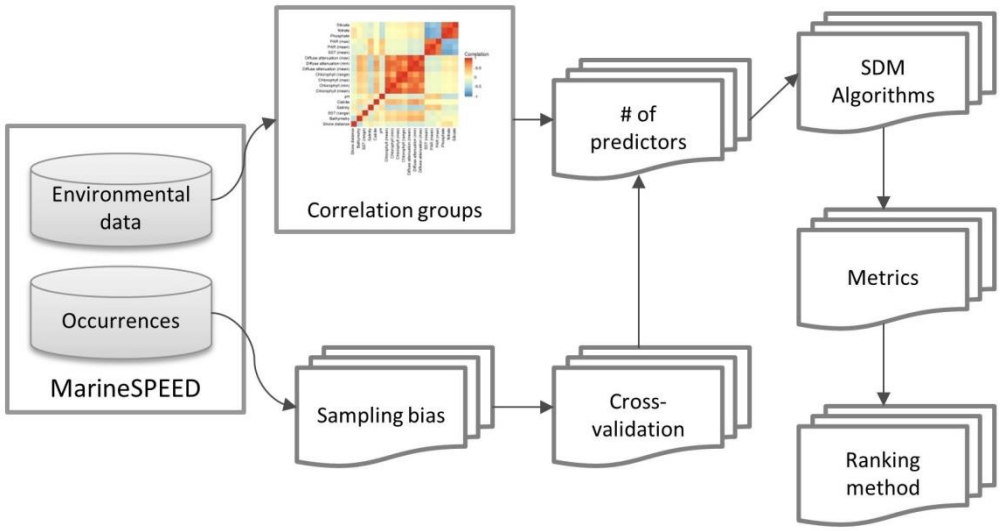


Figure 1. Overview of the predictor selection analysis and the different steps where variations were introduced. Starting from 19 environmental predictors, from Bio-ORACLE and MARSPEC, correlation groups where created. From this all possible predictor combinations were generated for models with three, four and seven predictors. After optional sample selection bias mitigation, occurrence records and background points were split in random or spatial cross-validation folds. SDMs were build using four algorithms (random forests, MaxEnt, generalized linear models and Bioclim) and evaluated using the area under the curve of the receiver operating characteristic (AUC) and the point-biserial correlation (COR). Predictors were ranked based on the performance of the models they were included in.

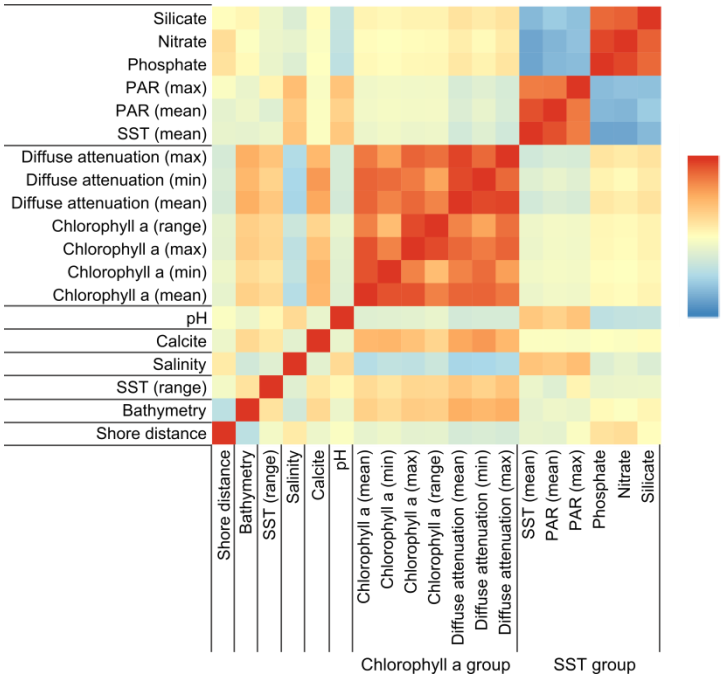


Figure 2. Correlation matrix for all environmental predictors considered for the predictor selection analysis, sorted by correlation group. Note that for creating the correlation groups, predictors are grouped when the absolute correlations between two or more members of a correlation group is more than 0.70. Red indicates a high positive correlation, yellow no correlation and blue a high negative correlation.

SDMs were fitted using four commonly used algorithms: Bioclim (Booth et al., 2014), Generalized Linear Model (GLM), Maximum Entropy modelling (Maxent, Phillips et al. 2004) and Random Forests (RF, Breiman 2001). We used the *dismo* package (Hijmans et al., 2016) in R for fitting Bioclim and MaxEnt models and the R package *randomForest* (Liaw & Wiener, 2002). For all algorithms the default settings were used and GLMs were run with only linear features.

Three variations of sample selection bias correction were performed: 1) no correction, 2) spatial thinning (50 km) with the R package *spThin* (Aiello-Lammens et al., 2015) and a target-group background (Phillips et al., 2009). Performance of the models was evaluated using random as well as spatial disc-based cross-validation. In total six million models were fitted and evaluated using the area under the receiver operating characteristics (ROC) curve (AUC) (Hanley & McNeil, 1982), and the point-biserial correlation (COR) (Zheng & Agresti, 2000; Elith et al., 2006) on the UGent High Performance Cluster.

Predictors where ranked for each model setup, evaluation metric and species combination by ranking the mean or median performance of all models a predictor was used in and by using the Rank Centrality algorithm (Negahban et al., 2017). Rank

Centrality is an iterative algorithm for rank aggregation using pairwise-wise comparisons.

Results

Benchmark data set

The MarineSPEED benchmark dataset is composed of 514 species with an original total of two million distribution records which have been filtered down on a 25 km² grid to nearly nine hundred thousand records. On a species level the median number of filtered distribution records is 506 with a minimum of 52 and a maximum of 45,469. An overview of the information on the species is available in Appendix S2.

A total of 18 different phyla are included in MarineSPEED (Fig. 3), with as most represented phyla: Chordata (245 species), Mollusca (62 species), Echinodermata (38 species), Arthropoda (36 species) and Annelida (32 species). The phylum Chordata is mostly represented by the class Actinopterygii (184 species), and to a lesser extent Elasmobranchii (20 species) and Mammalia (18 species). Marine primary producers, various groups of algae and seagrasses, are represented by 49 species from 5 phyla. When classifying species into functional groups we see that 395 species are associated with the seafloor (benthos), while 87 species are free swimming (nekton) and 32 species are planktonic. While we aimed to select species from different parts of the world a bias towards a few well-researched areas (e.g. the North-Atlantic and Australia) was unavoidable (Fig. 4). Likewise, coastal areas (442 species) are overrepresented compared to open ocean habitats (72 species). On a latitudinal scale, temperate regions are the most represented with 173 species. 91 species only occur in the tropics and 11 species in the polar regions. When considering the sample selection bias criterion we see that 59 species have a very low degree of sample selection bias (value 1), that most species have value 2 (103 species), 3 (156 species) or 4 (178). Only 18 species were assessed as having a very high degree of sample selection bias.

The predefined spatial cross-validation splits all considerably increase the distance between test points and their nearest training point as compared to random splits (Fig. S1 in Appendix S3). Examples of the various cross-validation strategies are visualised for *Didemnum maculosum* Milne Edwards and *Polycarpa aurata* Quoy & Gaimard in Fig. S2 and S3, respectively in Appendix S3.

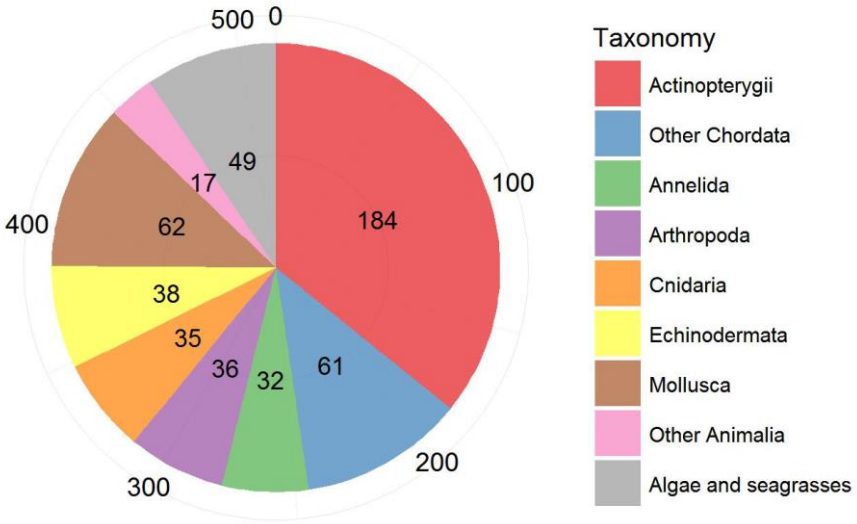


Figure 3. Taxonomic composition of the MarineSPEED dataset on level kingdom, phylum or class. For the kingdom Animalia the most abundant phylum Chordata was split up into the Actinopterygii and other Chordata, the kingdom Plantae was left as one whole and labelled as algae and seagrasses. Numbers represent the number of species in each taxonomic group.

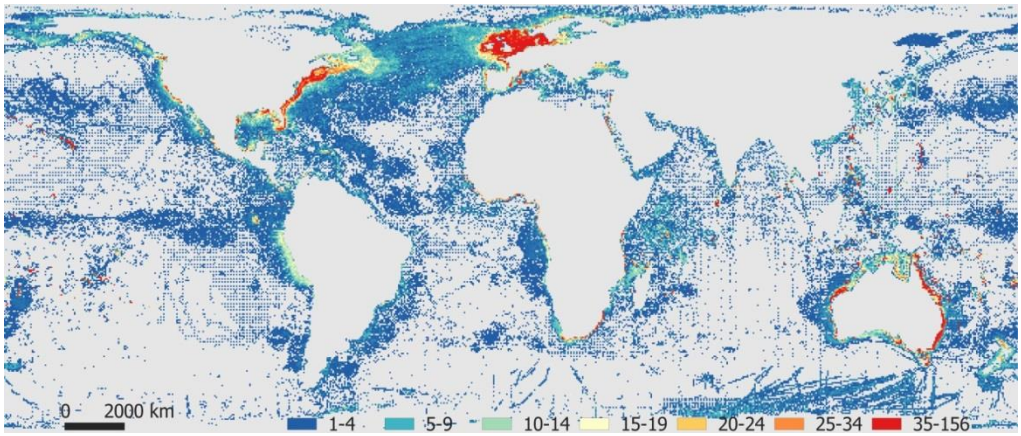


Figure 4. Map of the number of species occurring in each cell of an equal-area grid with a per cell area of 25 km² (Behrmann cylindrical equal-area projection).

Predictor relevance

A first set of analyses exploring the selection of relevant predictors (Fig. 5), highlights the importance of mean sea surface temperature (SST (mean)) as the most relevant predictor of the species distributions in the MarineSPEED benchmark dataset. This result appears robust regardless of modelling algorithms, sample selection bias correction, cross-validation, number of predictors, evaluation metrics and ranking

methods. At the other end of the spectrum, calcite is apparently irrelevant as a predictor for most of the species distributions. As for the other predictors, however, there is substantial variation across species and modelling parameters.

Among the different algorithms, GLMs with linear features caused the most variation in the predictor top 5 rankings with a particularly strong effect on SST (mean) with a minimal decrease of 28% in the median percentage of species with SST (mean) in the top 5 ranking (Table 1). Conversely in GLMs bathymetry was selected at least 26% more. The difference between the two evaluation metrics AUC and COR on the other hand was fairly limited with salinity displaying the largest difference. Finally the ranking method showed very small differences between the mean and median ranking algorithm. The rank centrality algorithm consistently ranked the predictors from the "Chlorophyll a group" as less relevant, while increasing the ranking of salinity (+16%) bathymetry (+15%), pH (+13%) and shore distance (+13%).

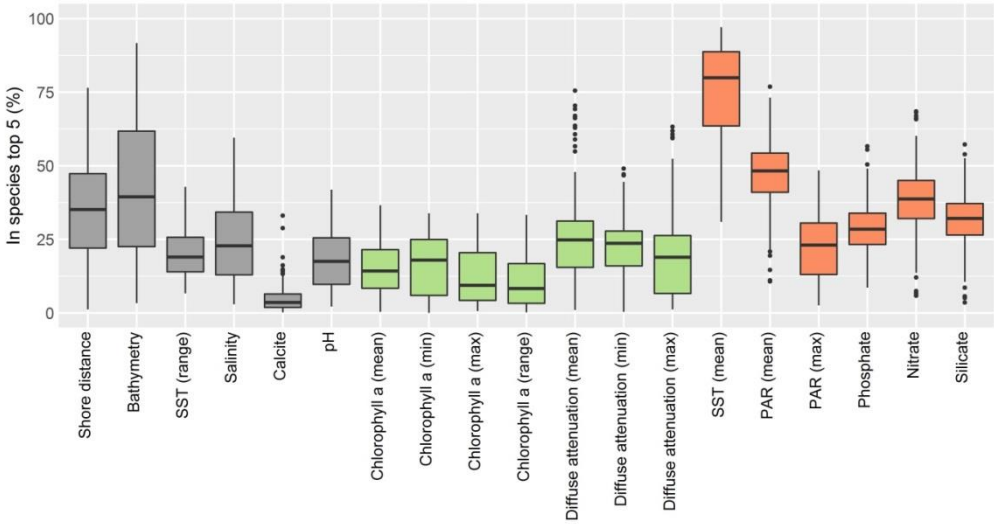


Figure 5. Percentage of species a predictor has a top 5 ranking in the different model setups. In grey are the predictors that form a correlation group on their own, in green the predictors from the "Chlorophyll a group" and in red the predictors from the "SST group". The results are aggregated from all possible variations. For a detailed view on the different dimensions of the variations we refer to tables 1 to 3, and to the following plots in Appendix S3: modelling algorithms (Fig. S4), evaluation metrics (Fig. S5), ranking methods (Fig. S6), cross-validation strategies (Fig. S7), predictor counts (Fig. S8), sampling bias mitigation methods (Fig. S9), cross-validation folds (Fig. S10) and taxonomic groups (Fig. S11).

When comparing the results of CV splitting strategies, number of predictors, sampling bias mitigation and fold number (Table 2), we can conclude that the number of predictors allowed in the model has the largest effect. Increasing the number of allowed predictors from 3 to 7 causes a decline in the relevance of bathymetry (-31%) and shore distance (-26%) while increasing the relevance of PAR (max) (+17%), diffuse attenuation (max) (+14%) and chlorophyll a (max and range) (+13%). The second largest effect is caused by using a target-group background in order to mitigate the effect of sampling bias on SDMs with a decrease of 25% for bathymetry and 15% for shore distance and an increase of 12% for nitrate. When using the disc-based CV strategy the relevance of SST (mean) and salinity decreased with 19 and 10%, respectively. Using the second fold instead of the first fold, which was only performed for the random CV strategy, only yielded small differences in the top 5 predictors of the species.

While the relevance of most predictors, is similar across taxonomic groups, some predictors exhibit large differences (Table 3). This is especially the case for shore distance, bathymetry and SST (range) with differences between the minimum and maximum of 55, 40 and 33%, respectively. Despite these overall patterns in the median ranking values we see that the spread of the predictor relevance within taxonomic groups is large (Fig. S11).

Table 4 presents the results related to the different traits of the species: functional group, neritic versus oceanic zone, ecoregion and sampling bias. Regarding the functional group some clear trends are visible whereby shore distance, bathymetry and to a lesser extent PAR (mean) are comparatively more relevant predictors for benthic species distributions, less relevant for nekton and least relevant for plankton. For mean and minimum diffuse attenuation we notice an inverse trend with a higher relevance for plankton in comparison to nekton and benthos. With respect to the zone trait we see that shore distance (-21%) and bathymetry (-14%) are less relevant for oceanic species, while phosphate (+15%), nitrate (+13%) and silicate (+15%) are more relevant. The results from the ecoregion trait show clear differences in predictor relevance for multiple predictors. For some predictors such as SST (range), nitrate and phosphate the relevance for temperate species clearly deviates from that for polar and tropical species. The predictor relevance for the different levels of sampling bias are all very similar. For boxplots of the relevance of the predictors for the different variations in model setup, taxonomic groups and traits we refer to Figs. S4 to S15 in Appendix S3.

Table 1. Median percentage of species for which a predictor has a top 5 ranking for the different setup variations that have been calculated for all models. First column shows the results for all models, the next four columns show the results for the different modelling algorithms, the next two columns show the breakdown for the evaluation metrics used. The last three columns show the results for the ranking methods.

			Algorithm				Metric		Ranking method		
Group	Predictor	**All**	Bioclim	GLM	MaxEnt	RF	AUC	COR	Centrality	Mean	Median
	Shore distance	**35**	29	22	39	40	36	34	44	31	27
	Bathymetry	**39**	45	71	36	19	40	37	52	37	33
	SST (range)	**19**	14	24	19	18	18	19	26	16	16
	Salinity	**23**	16	15	25	37	18	26	33	17	16
	Calcite	**4**	4	5	3	3	3	4	6	2	3
	pH	**18**	8	24	14	23	17	18	26	12	13
Chlorophyll a group	Chlorophyll a (mean)	**14**	18	8	14	17	15	13	9	16	18
	Chlorophyll a (min)	**18**	22	4	21	21	17	18	6	22	22
	Chlorophyll a (max)	**9**	15	6	11	15	10	9	5	17	19
	Chlorophyll a (range)	**8**	11	7	9	13	8	9	3	13	15
	Diffuse attenuation (mean)	**25**	21	44	24	24	23	26	10	27	27
	Diffuse attenuation (min)	**24**	22	30	22	21	22	23	9	25	25
	Diffuse attenuation (max)	**19**	12	37	10	16	18	19	7	23	23
SST Group	SST (mean)	**80**	79	51	89	86	79	78	79	79	78
	PAR (mean)	**48**	53	49	48	41	46	49	51	46	46
	PAR (max)	**23**	22	30	20	15	20	24	26	17	22
	Phosphate	**28**	32	23	27	32	29	27	33	26	26
	Nitrate	**39**	41	31	41	44	41	33	41	38	37
	Silicate	**32**	27	29	32	36	32	31	36	29	31

Table 2. Median percentage of species for which a predictor has a top 5 ranking for the different setup variations that have been calculated for a subset of the models. In this table only results from setups that have been done for both options are shown. First column shows the results for all models, followed by the results for the 5-fold random and disc-based spatial cross-validation splitting strategies, the breakdown for the number of predictors used in the models, the impact of sampling bias mitigation techniques and the results for the first and the second fold.

			CV splitting strategy		Predictor count			Sampling bias mitigation			Fold number	
Group	Predictor	**All**	Disc	Random	3	4	7	None	spThin	Target-group	1	2
	Shore distance	**35**	35	30	56	55	30	30	27	12	30	35
	Bathymetry	**39**	42	34	65	62	34	34	33	8	34	37
	SST (range)	**19**	15	21	19	24	21	21	18	18	21	11
	Salinity	**23**	13	23	22	28	23	23	23	28	23	20
	Calcite	**4**	9	3	3	3	3	3	3	2	3	3
	pH	**18**	11	17	17	17	17	17	16	27	17	16
Chlorophyll a group	Chlorophyll a (mean)	**14**	15	18	12	12	18	18	15	22	18	19
	Chlorophyll a (min)	**18**	21	17	18	15	17	17	19	16	17	17
	Chlorophyll a (max)	**9**	16	16	3	4	16	16	17	19	16	14
	Chlorophyll a (range)	**8**	17	15	2	4	15	15	15	14	15	12
	Diffuse attenuation (mean)	**25**	18	26	24	24	26	26	26	28	26	27
	Diffuse attenuation (min)	**24**	24	24	25	21	24	24	25	19	24	24
	Diffuse attenuation (max)	**19**	18	20	6	8	20	20	21	25	20	22
SST Group	SST (mean)	**80**	59	78	85	84	78	78	80	85	78	76
	PAR (mean)	**48**	46	50	37	47	50	50	51	59	50	49
	PAR (max)	**23**	34	25	8	12	25	25	25	25	25	23
	Phosphate	**28**	32	26	28	27	26	26	27	28	26	30
	Nitrate	**39**	36	33	42	38	33	33	34	46	33	44
	Silicate	**32**	36	35	29	29	35	35	32	29	35	29

Table 3. Median percentage of species for which a predictor has a top 5 ranking for the different setup variations that have been calculated for all models and for some taxonomic groups. Within the class Chordata and within the kingdom Animalia taxonomic groups with few species were left out of this comparison.

			Chordata	Other Animalia					Plantae
Group	Predictor	**All**	Actinopterygii	Annelida	Arthropoda	Cnidaria	Echinodermata	Mollusca	Algae and seagrasses
	Shore distance	**35**	42	16	11	66	32	29	44
	Bathymetry	**39**	49	42	33	54	51	31	14
	SST (range)	**19**	14	42	36	9	13	19	14
	Salinity	**23**	18	16	19	11	21	25	31
	Calcite	**4**	2	3	6	3	8	3	4
	pH	**18**	19	6	11	11	13	21	18
Chlorophyll a group	Chlorophyll a (mean)	**14**	11	9	17	9	16	15	18
	Chlorophyll a (min)	**18**	15	16	19	9	16	15	20
	Chlorophyll a (max)	**9**	9	5	11	6	13	10	10
	Chlorophyll a (range)	**8**	8	3	8	6	11	10	8
	Diffuse attenuation (mean)	**25**	17	31	33	11	18	27	35
	Diffuse attenuation (min)	**24**	17	31	25	9	21	23	29
	Diffuse attenuation (max)	**19**	19	9	17	14	21	18	18
SST Group	SST (mean)	**80**	81	81	72	83	71	69	76
	PAR (mean)	**48**	52	41	42	57	45	47	33
	PAR (max)	**23**	18	28	33	9	21	24	18
	Phosphate	**28**	29	19	33	29	26	26	23
	Nitrate	**39**	43	22	36	47	34	35	24
	Silicate	**32**	25	41	36	14	29	35	39

Table 4. Median percentage of species for which a predictor has a top 5 ranking for the different setup variations that have been calculated for all models and traits. For the functional group trait, benthos includes all seafloor associated species, including demersal and reef-associated species; nekton includes all actively swimming pelagic species and plankton are all species unable to swim against a current. The neritic and oceanic zones were defined based on the ecoregion classification by Spalding (2007) whereby species having 5% or more of their distribution records outside of ecoregions are classified as oceanic. Species are a member of an ecoregion when at least 5% of its distribution records are situated in a polar, temperate or tropical ecoregion. Sampling bias was visually assessed from 1 (low bias) to 5 (high bias).

			Functional group			Zone		Ecoregion			Sampling bias				
Group	Predictor	**All**	Benthos	Nekton	Plankton	Neritic	Oceanic	Polar	Temperate	Tropical	1	2	3	4	5
	Shore distance	**35**	39	24	13	38	17	13	25	49	29	28	42	35	28
	Bathymetry	**39**	44	26	13	40	26	39	25	60	22	30	47	44	22
	SST (range)	**19**	17	22	28	19	19	13	28	5	19	28	15	18	17
	Salinity	**23**	20	24	22	22	18	26	27	14	31	28	17	17	28
	Calcite	**4**	3	2	3	3	1	0	3	2	5	4	3	3	6
	pH	**18**	17	17	6	19	7	4	18	14	26	22	13	13	14
Chlorophyll a group	Chlorophyll (mean)	**14**	12	16	19	13	17	17	17	10	14	14	13	13	17
	Chlorophyll (min)	**18**	16	19	17	17	17	13	20	11	20	20	14	15	22
	Chlorophyll (max)	**9**	8	11	9	9	10	4	10	8	12	9	8	10	11
	Chlorophyll (range)	**8**	8	8	9	8	10	4	8	8	10	8	7	8	11
	Diffuse attenuation (mean)	**25**	22	30	44	24	25	30	33	12	24	28	21	24	28
	Diffuse attenuation (min)	**24**	21	27	34	23	24	22	31	10	24	24	21	22	22
	Diffuse attenuation (max)	**19**	18	16	16	19	15	13	16	20	17	15	19	19	22
SST Group	SST (mean)	**80**	79	77	78	79	77	74	74	86	63	74	85	81	72
	PAR (mean)	**48**	49	45	34	48	44	26	42	56	42	46	53	46	42
	PAR (max)	**23**	21	29	19	22	19	17	25	14	25	26	19	20	22
	Phosphate	**28**	27	25	34	25	40	48	21	34	25	22	29	29	33
	Nitrate	**39**	39	33	31	36	49	57	27	49	30	28	41	43	28
	Silicate	**32**	28	39	41	28	43	52	38	16	37	36	23	31	33

Data access

While distribution maps for all species can be consulted and all data is downloadable in an R Shiny interface (Chang et al., 2016) at <http://marinespeed.org>, we opted to also create the *marinespeed* R package allowing for an easy usage of the data (Table 4). The first step, after installation from CRAN and loading the library, is to run the function 'list_species' which returns the scientific names and WoRMS identifiers for all species. Additional information on the taxonomy, sampling bias estimate and latitudinal zones can be viewed using the 'species_info' function. In order to run a function for all species either the 'lapply_species' or the 'lapply_species_kfold' function can be used. Alternatively, if you only need data for specific species, the 'get_occurrences' and 'get_fold_data' methods can be used. On top of this other lower level functions for loading background data and creating cross-validation splits are also available.

Table 4. Overview of the most important functions in the *marinespeed* R package. Lower level functions for accessing occurrences, background data and creating cross-validation folds are also available.

Function	Description
list_species	Get the list of scientific names and WoRMS identifiers for all species.
species_info	Additional species information.
lapply_species	Execute a function for all distribution records for multiple species.
lapply_kfold_species	Execute a function for one or more pre-made CV folds for multiple species.

Discussion

Species distribution modelling is widely used to identify areas that are ecologically suitable for the presence of species under past, current and future climates. Most studies concentrate, however, on terrestrial environments, while marine species distribution modelling kicked off comparatively late (Robinson et al., 2011). A direct consequence of the relative scarcity of marine SDM studies is that most of the methodological progress in SDM is biased towards terrestrial studies, despite marine environments being significantly different with respect to the ecological factors that control distributions and their spatio-temporal variation. These differences raise questions with respect to the environmental predictor relevance and the effects of model algorithms and settings on predictor relevance. By fitting presence-only SDMs for all combinations of predictors from different correlation groups, we assessed the

predictor relevance and the variation therein for marine species distributions. To this end, we created a benchmark dataset (MarineSPEED) which bundles marine species distributions of 514 taxa and associated environmental variables.

Relevant predictors

SST (mean) is the most relevant predictor of global marine species distributions, regardless of model algorithms and parameter settings. This result confirms the importance of temperature for species distributions identified in the meta-analysis by Bradie & Leung (2016) and its importance for the distribution of birds (Barbet-Massin & Jetz, 2014). Moreover the importance of SST as a predictor in marine ecology was previously confirmed for marine species richness (Tittensor et al., 2010) and biogeographic structure of marine benthic fauna (Belanger et al., 2012). While bathymetry and shore distance also appear to be very relevant, there is considerable variance in the results, which might be because they are distal environmental predictors (Austin, 2002). In contrast to previous results (Nyström Sandman et al., 2013; Bradie & Leung, 2016) bathymetry was not the most important predictor, which can be explained by the global scale of our study. The importance of bathymetry has been shown to decrease with increased geographical scale (Nyström Sandman et al., 2013). Moreover the relevance of bathymetry is strongly linked to the species taxonomy (see Table 3 and 4 and Fig. S11-S14). At the other end of the spectrum, calcite is rarely selected as a meaningful predictor. The irrelevance of calcite is consistent with the fact that only one study in the meta-analysis by Bradie & Leung (2016) used calcite as a predictor. The remaining predictors are on average less part of the best scoring models, reflecting an overall reduced relevance toward predicting species distributions.

Despite this general trend the variance in predictor relevance is relatively high across model algorithms and settings.

The high variance when using different modelling algorithms is consistent with the results by Bucklin et al. (2015) who also demonstrated a significant interaction between predictor set and modelling algorithm. Especially predictor selection under GLM deviates from the other algorithms. Linear GLM-based models do not capture the relevance of SST (mean) very well. The lower relevance of SST in GLM models indicates that the global distribution of marine species is inadequately modelled by a linear relationship. Potentially, this effect can be mitigated by including polynomial features, an option which was not explored in the current analyses. In MaxEnt, with automatic selection of feature complexity and therefore yielding complex models, the relevance of SST (mean) is consistently high and displaying hardly any variation.

We expect that decreasing the complexity of the features fitted by MaxEnt will result in models more similar to GLM-based models. As for the other three algorithms, predictor selection seems to be largely consistent, echoing results of Barbet-Massin & Jetz (2014).

We also compared the predictor relevance under two different evaluation measures, AUC and COR, respectively. Although AUC, as an absolute measure for model performance, has been criticized earlier (Lobo et al., 2010) its use is warranted here as we only compared relative AUC values and only modelled in a fixed geographical extent. Both AUC, which measures the ability to discern presences from background data, and COR, which provides a measure for the calibration of the model showed very similar predictor rankings. This similarity is indicative for the generalizability of the results across model evaluation metrics.

Likewise, for most predictors the ranking method used did not affect the predictor relevance. The rank centrality method consistently gave a lower ranking to all predictors from the "Chlorophyll a group". As ranking from pairwise comparisons is an active research field, a future study comparing the rank centrality algorithm with other recent ranking methods such as spectral ranking (Fogel et al., 2016), sync rank (Cucuringu, 2016) and Microsoft's TrueSkill method (Herbrich et al., 2006) could lead to additional insights on the impact of the ranking algorithm on the predictor relevance.

The impact of cross-validation strategies was assessed by using spatial disc-based and random sampling of training and testing sets. Using a spatial instead of a random data splitting strategy in combination with the removal of spatial sorting bias resulted in a lower relevance of SST (mean). This can be attributed to two different factors: (1) extrapolation and (2) scale effects. Firstly, the spatial data splits sometimes causes a restriction in the predictor space, which leads to extrapolation (Roberts et al., 2016). With SST being in general the most relevant, extrapolation outside of its range will lead to low evaluation scores and therefore a lower ranking. On the other hand, due to the pairwise selection of test pseudo-absences at a similar distance to the test points as the distance between the test points and their nearest training point, the mean distance to evaluation background points decreases causing a scale effect. These results confirm that SST is especially relevant on a global scale but less so on a smaller scale (Nyström Sandman et al., 2013).

Restricting the number of predictors included in a model directly influences the relevance of the predictors. For most marine species the relevance of bathymetry and shore distance diminishes when more predictors are included in the model.

These predictors are only distally related to the suitability of an environment for species distributions and therefore the potential choice of more proximate predictors will result in their lower relevance in predictor-rich models. Inversely predictors from the "Chlorophyll a group" are selected more, suggesting that if combined with some of the predictors from the other correlation groups they provide a better explanation of the species distribution then bathymetry and shore distance do.

Unlike the effect of spatial thinning, using a target-group background resulted in large differences in predictor relevance. As most of the species occurrence records are located along the coast, the target-group background, which is a subsample of it, is expected to have the same bias resulting in a lower relevance of shore distance and bathymetry. These results confirm the importance of background selection on SDMs (Chefaoui & Lobo 2008; Phillips et al. 2009; VanDerWal et al. 2009; Barbet-Massin et al. 2012; Acevedo et al. 2012; Smith 2013; Senay et al. 2013). It is therefore recommended to investigate the impact of alternative pseudo-absence selection methods in future studies. Note that in general it is advised to create a species specific target-group with occurrence records from the same sampling campaign(s) and/or from similar species, reflecting the sampling bias of the species modelled (Phillips et al., 2009).

In this study we explored the impact of several parameter settings on predictor selection, however the potential analyses are by no means exhaustive. For example the regularisation parameter and the complexity of the features in MaxEnt, the number of trees fitted in random forests and the usage of polynomial features in GLM were kept constant or were not explored. It is likely that applying species-specific tuning of the algorithms will not only impact model performance but also affect the predictor selection (Anderson & Gonzalez, 2011; Merow et al., 2014).

From a species perspective we noted that the taxonomy and the traits of a species have an influence on the relevance of predictors. The overarching pattern of predictor relevance holds up across traits, but some marked differences in predictor relevance were found for shore distance and bathymetry and to a lesser extent for diffuse attenuation, phosphate, nitrate and silicate. To some extent these differences are intuitive. For example, subdividing the taxa between oceanic and neritic species results in a higher relevance of shore distance for neritic species. Likewise, SST range is less relevant for tropical and polar species, because low and high latitudes typically exhibit very little annual sea surface temperature fluctuations

compared to mid-latitudes. Despite some pronounced differences across traits, trends for inorganic nutrients (nitrate, phosphate, silicate) are less easily explained.

Benchmark dataset

Inspired by the widespread use of benchmark datasets in machine learning and other computational fields we set out to create MarineSPEED. Although a series of papers was published using the same set of 226 terrestrial species (e.g. Elith et al. 2006; Guisan et al. 2007; Phillips et al. 2009; Hijmans 2012) most studies discussing new methods related to SDM use a small set of different species. Moreover while the resulting algorithm and methods are regularly made available through ready to use R packages or desktop programs, the species distribution records used in these studies often are not. With the release of MarineSPEED and its associated R package researchers can download all occurrences, background records and cross-validation data sets.

The marine character of the dataset is ideally suited for the study of methodological issues and parameterizations for distribution modelling of non-terrestrial species. This is necessary as the marine environment poses its own challenges for SDM (Kaschner et al., 2006; MacLeod et al., 2008; Dambach & Rödder, 2011; Robinson et al., 2011; Bentlage et al., 2013). Species distribution records from public databases contain a combination of opportunistic records and systematic sampling campaigns. They show large biases in amount and location of occurrences where the coastal areas are often more intensely sampled than offshore areas. The lower detectability of marine species in combination with the wide extent of the marine environment leads to false absences and a general lack of distribution records in comparison to the real world range extent of marine species. MacLeod et al. (2008) found that in contrast to the terrestrial environment, presence-absence methods don't perform better than presence-only methods in the marine environment. Although absences are rarely reported for marine species and not included in MarineSPEED, this study could be confirmed by using estimated absence data for species included in systematic surveys in OBIS (Coro et al. 2016).

Applications

Combining the marinespeed R package with one of the numerous SDM packages like *BIOMOD2*, *dismo*, *sdm* or *zoon*, other machine learning packages like *caret*, *gbm*, *randomForest* or *xgboost* and the general R ecosystem allows for numerous applications.

While several papers have compared the performance of SDM algorithms (e.g. Elith et al. 2006; Tsoar et al. 2007; Meynard & Quinn 2007; Liu et al. 2011; Lorena et al. 2011), new SDM modelling algorithms are regularly released (e.g. MaxLike (Royle et al., 2012), Plateau (Brewer et al., 2016), GRaF (Golding & Purse, 2016)). Consistent usage of MarineSPEED to explore the performance of modelling algorithms would allow for a direct comparison of the strengths and weaknesses of them. On top of this, SDM algorithms benefit from species-specific parameter settings (Anderson & Gonzalez, 2011; Merow et al., 2013; Shcheglovitova & Anderson, 2013) but useful ranges for the different parameters are unknown for these newer modelling algorithms.

Over the years, numerous studies have been published on methods for correcting sample selection bias (e.g. Dudík et al. 2005; Phillips et al. 2009; Boria et al. 2014; Varela et al. 2014; Barnes et al. 2014; Fernández & Nakamura 2015; Aiello-Lammens et al. 2015; Ranc et al. 2016) and selecting pseudo-absence records (e.g. Wisz & Guisan 2009; Lobo & Tognelli 2011; Barbet-Massin et al. 2012; Acevedo et al. 2012; Senay et al. 2013; Assis et al. 2015). Comparing these techniques with MarineSPEED can result in guidelines for sampling bias mitigation and pseudo-absence selection in the marine environment.

Next to the availability of marine species with environmental data and traits we expect that the *marinespeed* R package, with its implementation of cross-validation methods, to be a useful tool for SDM. Installation instructions, data downloads and species information can be found at <http://marinespeed.org/>.

Acknowledgements

The research was carried out with financial support from the ERANET INVASIVES project (EU FP7 SEAS-ERA/INVASIVES SD/ER/010) and financial, data & infrastructure support provided by VLIZ as part of the Flemish contribution to the LifeWatch ESFRI. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government – department EWI. We thank An Verfaillie, Taban Mestdag and Sara Martinez for contributing to the creation of the dataset. We further thank Bart Vanhoorne and Leen Vandepitte from WoRMS and Pieter Provoost from OBIS for providing support.

Data accessibility

The benchmark data can be downloaded from <http://marinespeed.org/>. The release version of the R package is on CRAN and the latest development version can be found at <https://github.com/lifewatch/marinespeed>.

Supporting information

Appendix S1

List of OBIS and GBIF datasets used for compiling MarineSPEED. Available at: http://www.phycology.ugent.be/research/marinespeed/MS_AppendixS1.docx.

Appendix S2

List of species included in MarineSPEED with their taxonomy, sampling bias, ecoregions and SST statistics. Available at: http://www.phycology.ugent.be/research/marinespeed/MS_AppendixS2.xlsx.

Appendix S3

Setup

Table S1. Overview of the different predictors used in the predictor selection analysis. The first column is the layer code used by the *sdmpredictors* R package to identify a predictor, the second column is the dataset the predictor was found in, the description column gives a short description of the predictor and the correlation groups column gives an indication of the correlation group a predictor belongs to.

Layer code	Dataset	Description	Correlation group
BO_chlomax	Bio-ORACLE	Chlorophyll a (maximum)	Chlorophyll a group
BO_chlomean	Bio-ORACLE	Chlorophyll a (mean)	Chlorophyll a group
BO_chlomin	Bio-ORACLE	Chlorophyll a (minimum)	Chlorophyll a group
BO_chlorange	Bio-ORACLE	Chlorophyll a (range)	Chlorophyll a group
BO_damax	Bio-ORACLE	Diffuse attenuation coefficient at 490 nm (maximum)	Chlorophyll a group
BO_damean	Bio-ORACLE	Diffuse attenuation coefficient at 490 nm (mean)	Chlorophyll a group
BO_damin	Bio-ORACLE	Diffuse attenuation coefficient at 490 nm (minimum)	Chlorophyll a group
BO_nitrate	Bio-ORACLE	Nitrate	SST group
BO_parmax	Bio-ORACLE	Photosynthetically available radiation (maximum)	SST group
BO_parmean	Bio-ORACLE	Photosynthetically available radiation (mean)	SST group
BO_phosphate	Bio-ORACLE	Phosphate	SST group
BO_silicate	Bio-ORACLE	Silicate	SST group
BO_sstmean	Bio-ORACLE	Sea surface temperature (mean)	SST group
BO_calcite	Bio-ORACLE	Calcite	Calcite
BO_ph	Bio-ORACLE	pH	pH
BO_salinity	Bio-ORACLE	Salinity	Salinity
BO_sstrange	Bio-ORACLE	Sea surface temperature (range)	SST range
MS_bathy_5m	MARSPEC	Bathymetry	Bathymetry group
MS_biogeo05_dist_shore_5m	MARSPEC	Distance to the shoreline	Shore distance

Table S2. Overview of all different setups for which models have been fitted for all combinations of predictors. Models for all species where build for all combinations of 3, 4 or 7 predictors using the random or disc-based splitting strategy to create the cross-validation (CV) data and the first or second fold from the 5-fold random cross-validation dataset. The last variation in setups is whether any and which sample selection bias correction method is used. For each predictor count we get a different total number of predictor combinations resulting in the calculation of a different number of models as models where fitted for all 514 species using 4 different SDM algorithms (bioclim, GLM, MaxEnt and random forests).

Predictor count	CV splitting strategy	Fold number	Sampling bias mitigation	Number of combinations	Number of models
3	Random	1	None	467	960,152
4	Random	1	None	905	1860,680
7	Disc-based	1	None	265	544,840
7	Random	1	None	265	544,840
7	Random	2	None	265	544,840
7	Random	1	spThin	265	544,840
7	Random	1	Targetgroup	265	544,840

Cross-validation splits

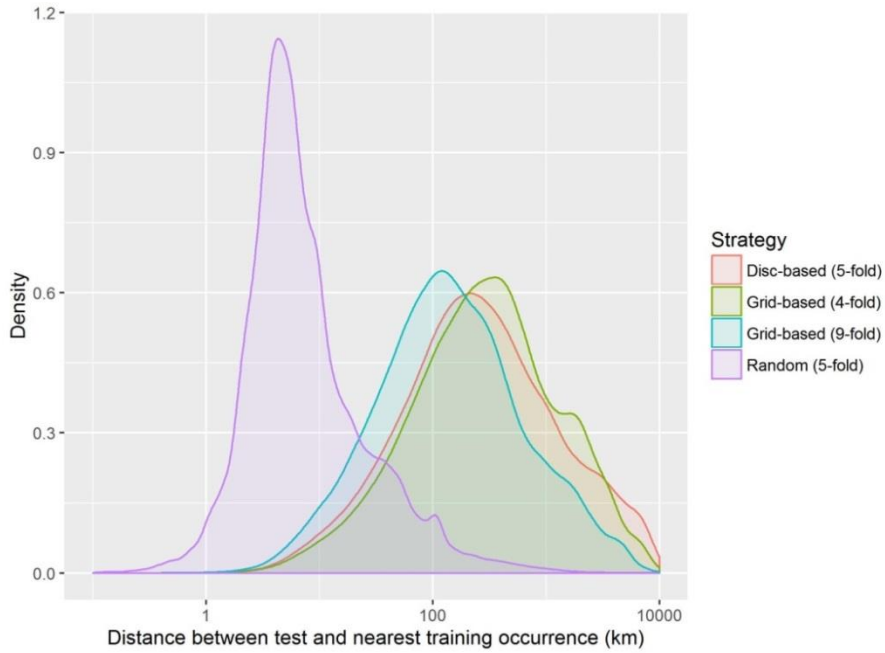
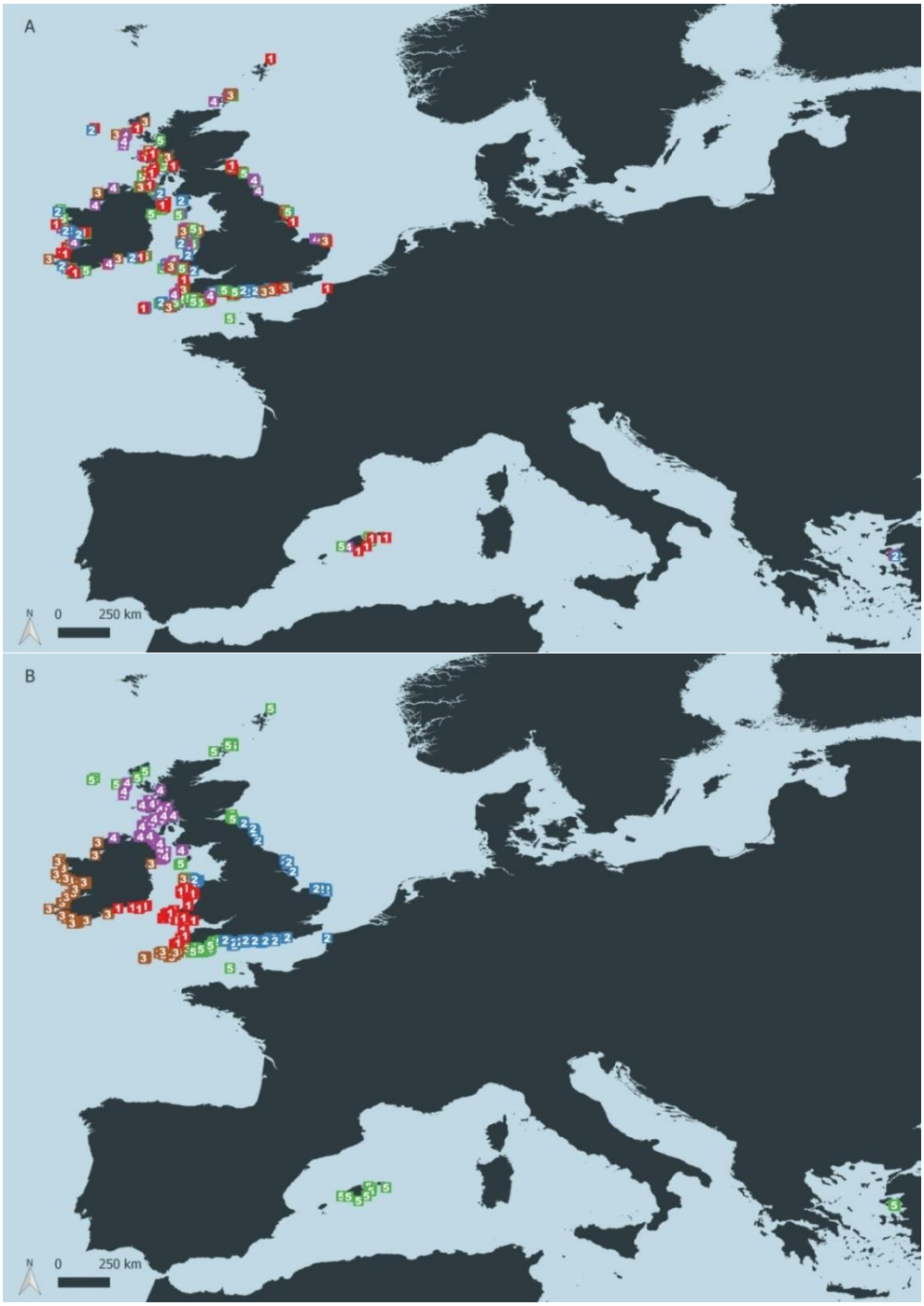


Figure S1. Density plot for the distance, on a log scale, between each test point and the nearest training occurrence for all folds of the four cross-validation splitting strategies with the 5-fold disc-based strategy in orange, the 4-fold grid-based strategy in green, the 9-fold grid-based strategy in blue and the 5-fold random strategy in purple.


A
N
0
250 km
B
N
0
250 km

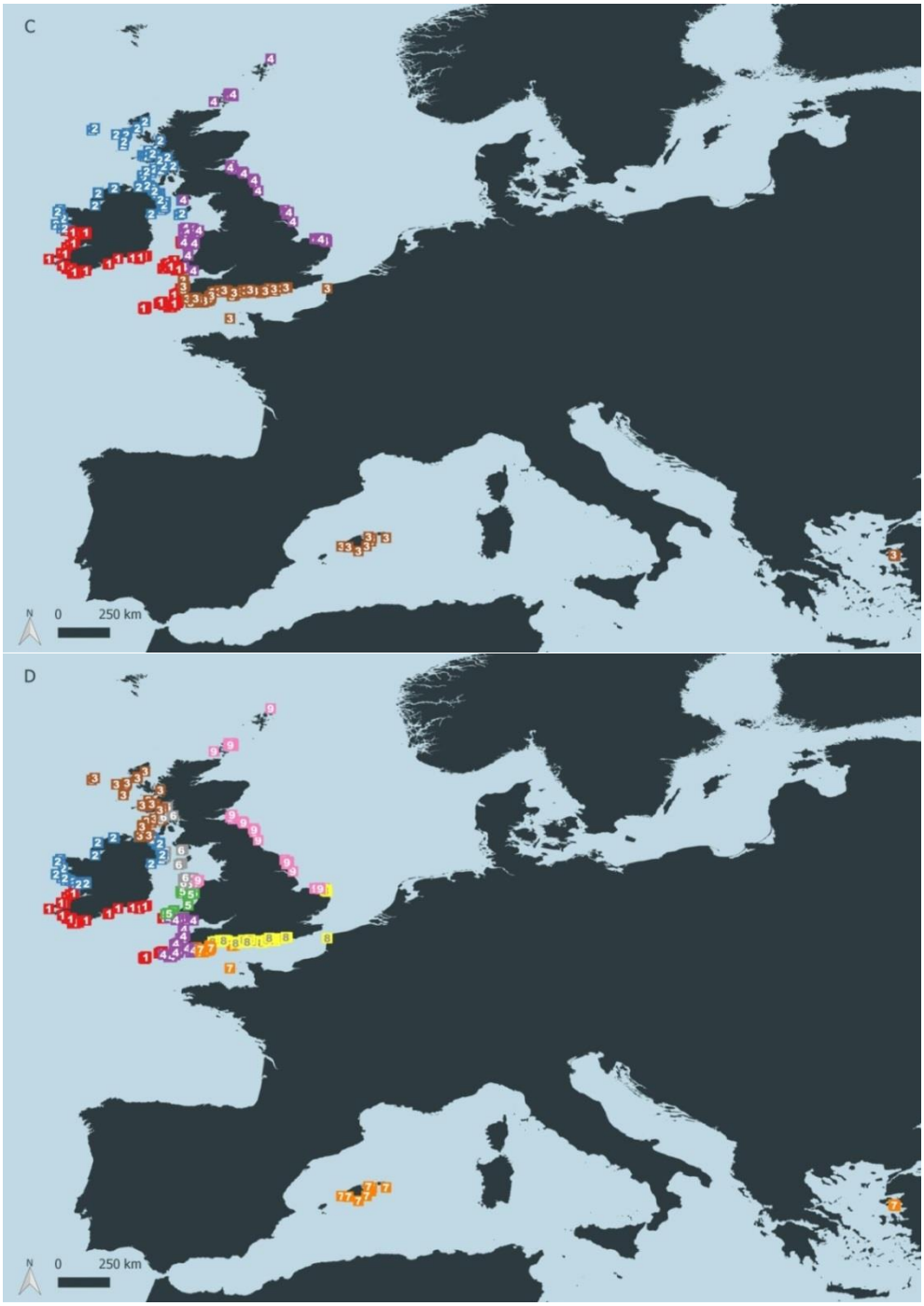
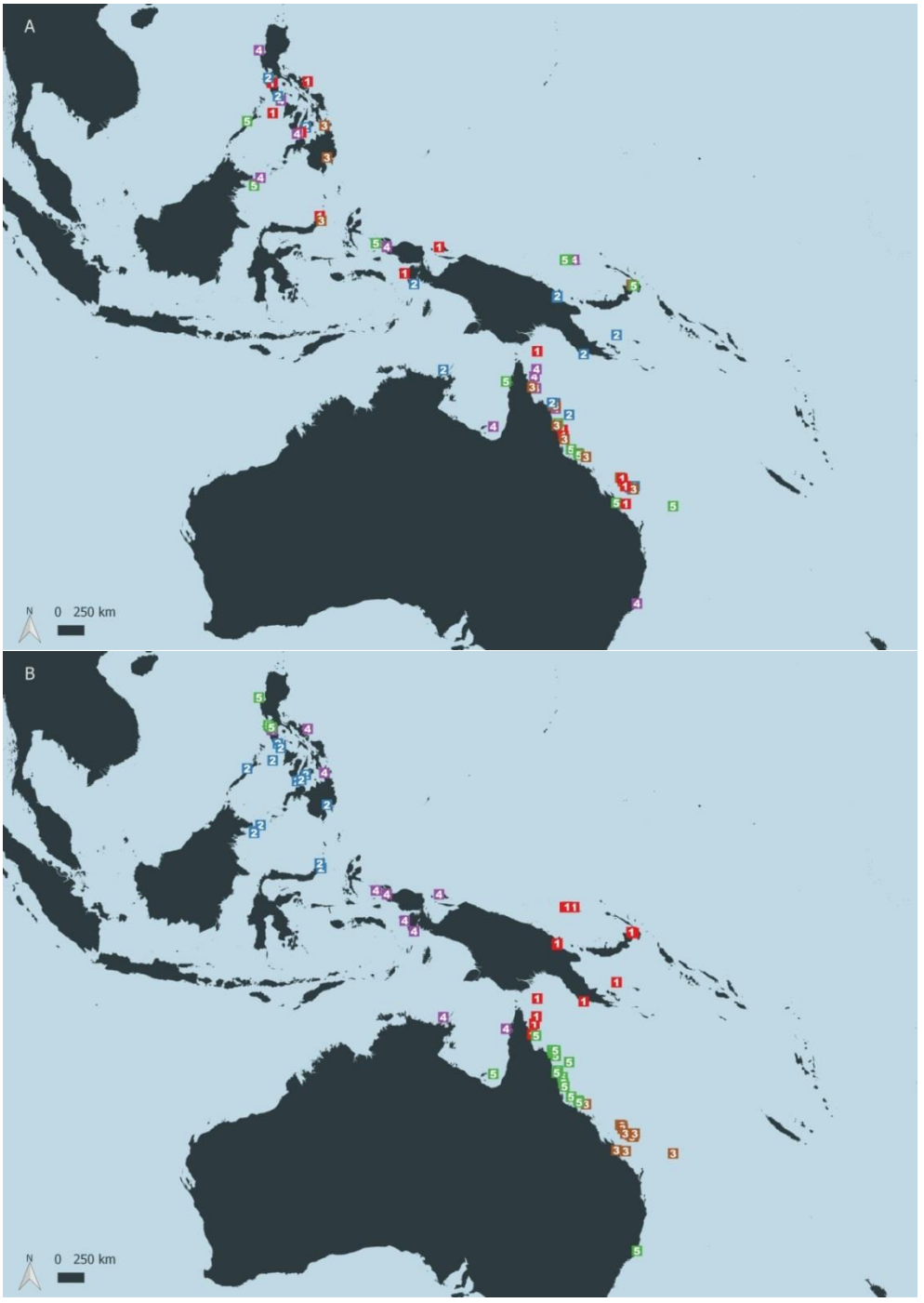


Figure S2. The cross-validation (CV) splits for the species *Didemnum maculosum* Milne Edwards using different methods: 5-fold random (A), 5-fold disc-based (B), 4-fold grid-based (C) and 9-fold grid-based (D). The different folds are numbered and coloured in the map (red=1, blue=2, brown=3, purple=4, green=5, grey=6, orange=7, yellow=8 and pink=9).


A
N
0 250 km
B
N
0 250 km

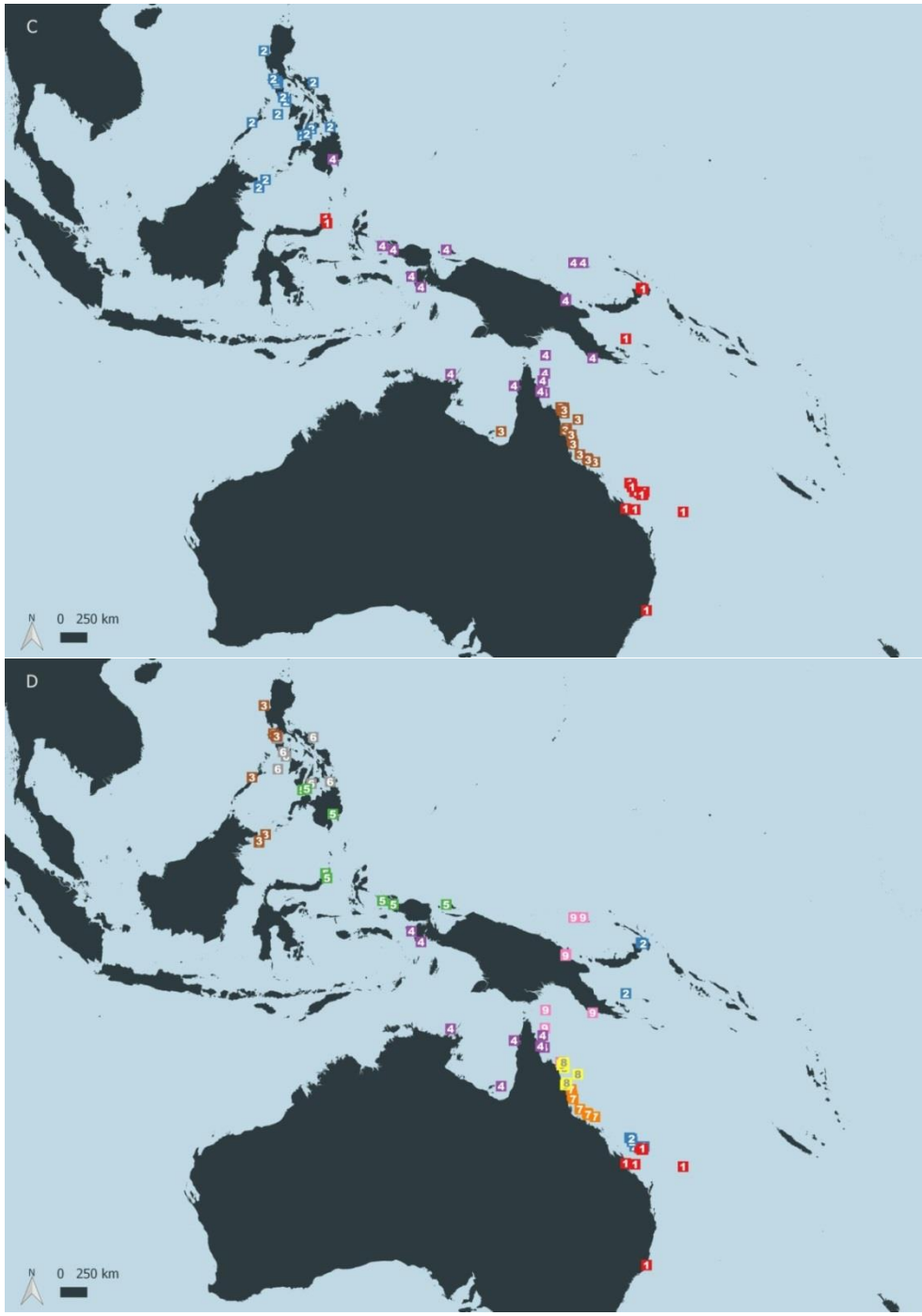


Figure S3. The cross-validation (CV) splits for the species *Polycarpa aurata* Quoy & Gaimard using different methods: 5-fold random (A), 5-fold disc-based (B), 4-fold grid-based (C) and 9-fold grid-based (D). The different folds are numbered and coloured in the map (red=1, blue=2, brown=3, purple=4, green=5, grey=6, orange=7, yellow=8 and pink=9).

Predictor relevance boxplots

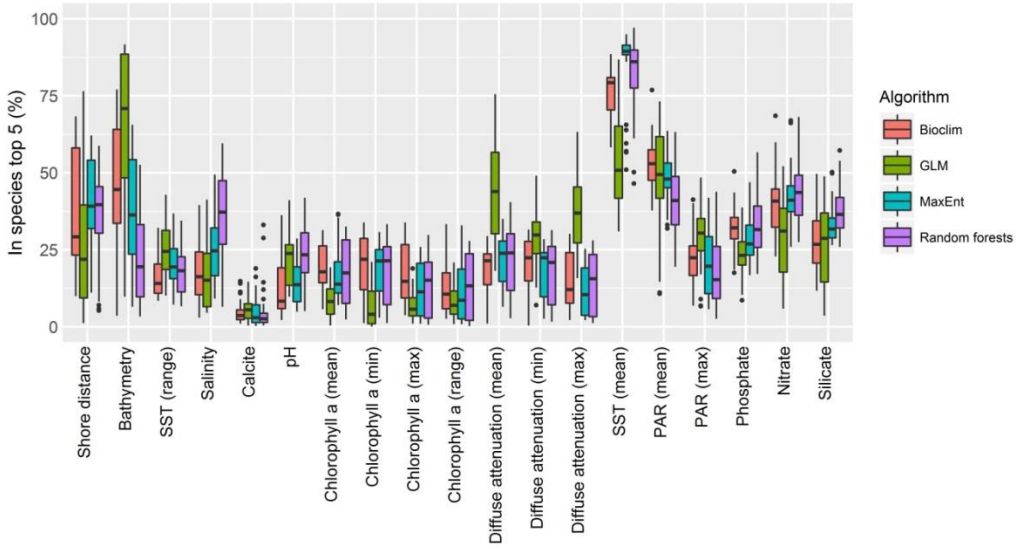


Figure S4. Percentage of species a predictor has a top 5 ranking in the different model setups for the different algorithms: bioclim (red), GLM (green), MaxEnt (blue) and random forests (purple).

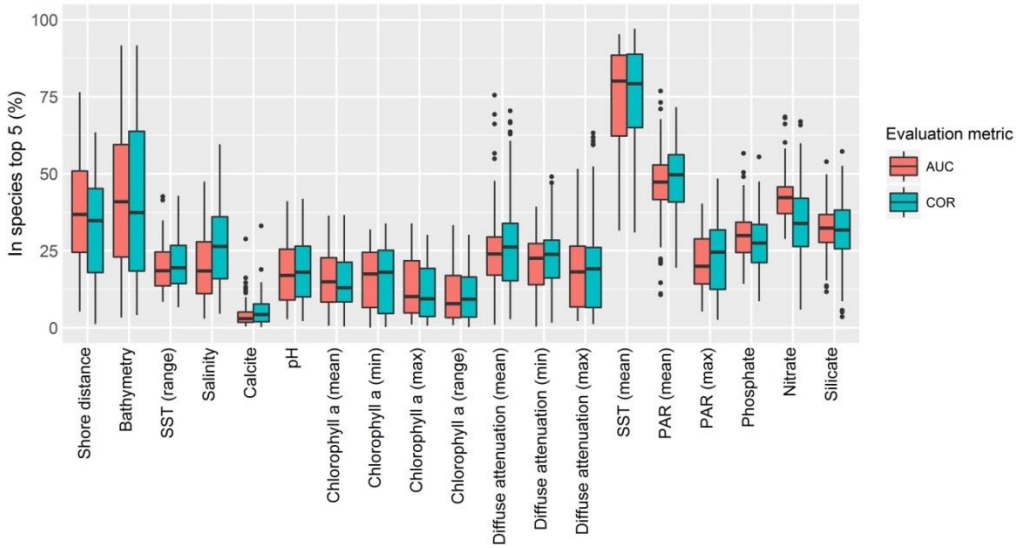


Figure S5. Percentage of species a predictor has a top 5 ranking in the different model setups for the two evaluation metrics: area under the receiver operating characteristic curve (AUC, red) and the point-biserial correlation (COR, blue).

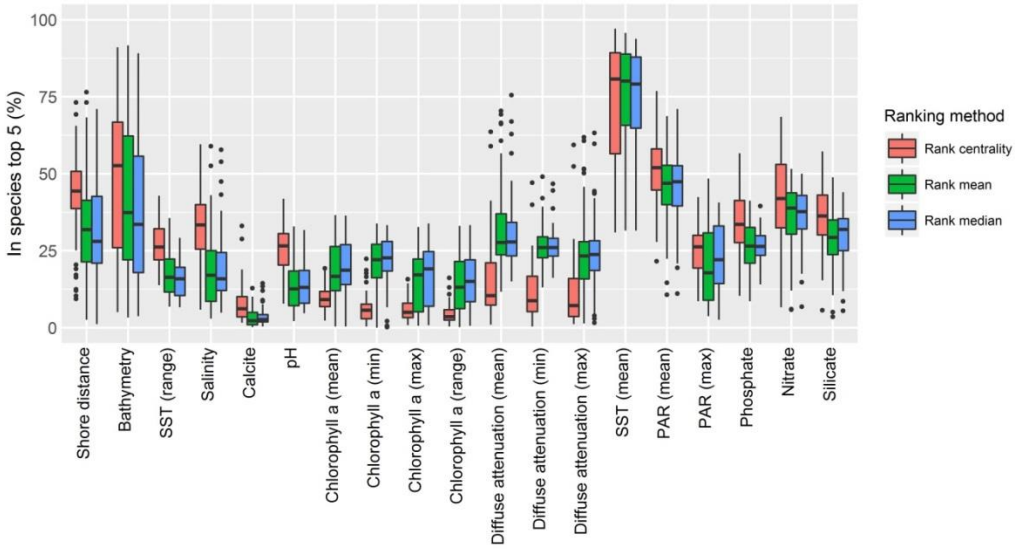


Figure S6. Percentage of species a predictor has a top 5 ranking in the different model setups for the three ranking methods: rank centrality (red), rank mean (green) and rank median (blue).

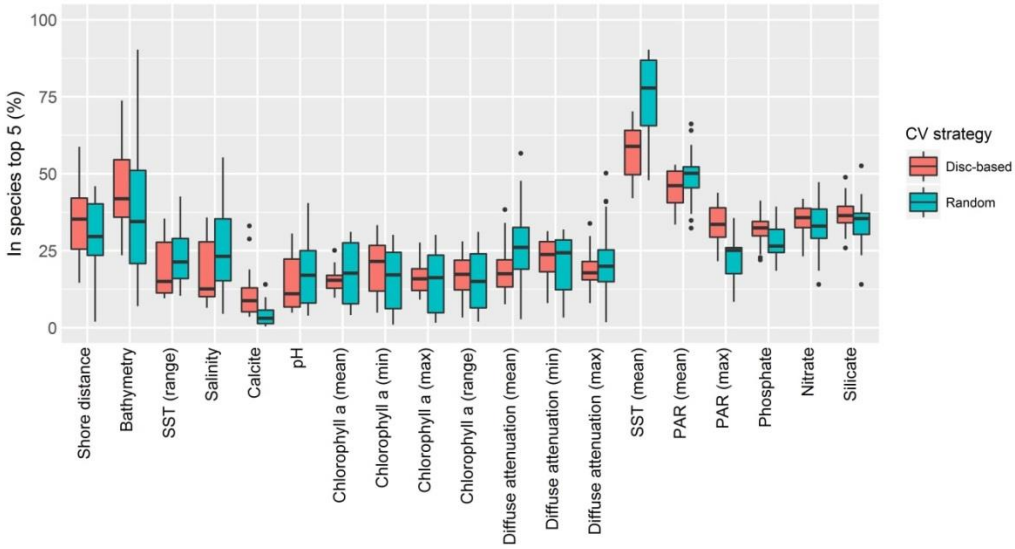


Figure S7. Percentage of species a predictor has a top 5 ranking in the different model setups for the two cross-validation (CV) strategies: disc-based CV (red) and random CV (blue). Note that only results for model setups that were run for both CV strategies are shown here.

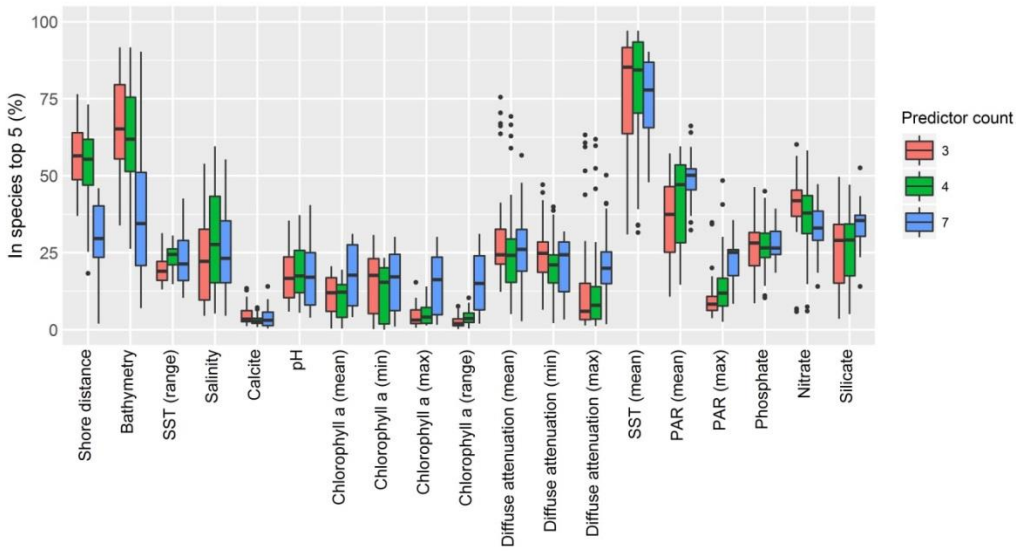


Figure S8. Percentage of species a predictor has a top 5 ranking in the different model setups for the different number of predictor counts: 3 (red), 4 (green), 7 (blue). Note that only results for model setups that were run for all three predictor counts are shown here.

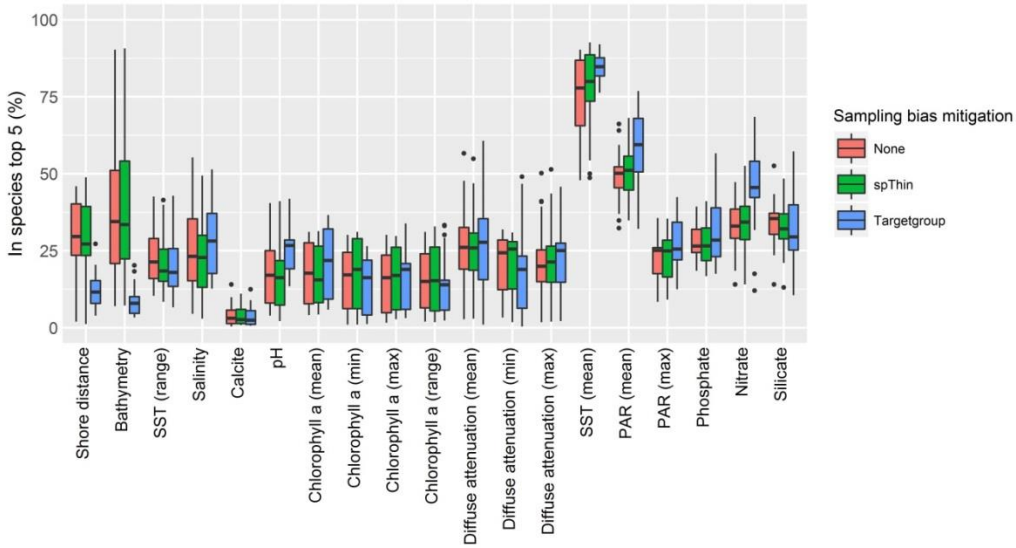


Figure S9. Percentage of species a predictor has a top 5 ranking in the different model setups for the different sampling bias mitigation methods: nothing (red), spatial thinning (spThin, green) and targetgroup background (blue). Note that only results for model setups that were run for all sampling bias mitigation methods are shown here.

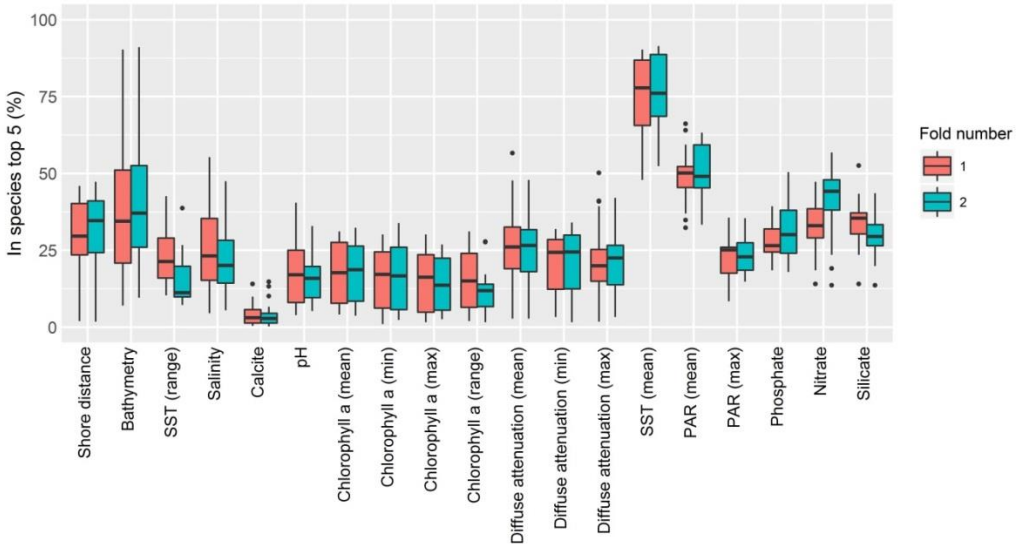


Figure S10. Percentage of species a predictor has a top 5 ranking in the different model setups for the two explored folds: 1 (red) and 2 (blue). Note that only results for model setups that were run for both folds are shown here.

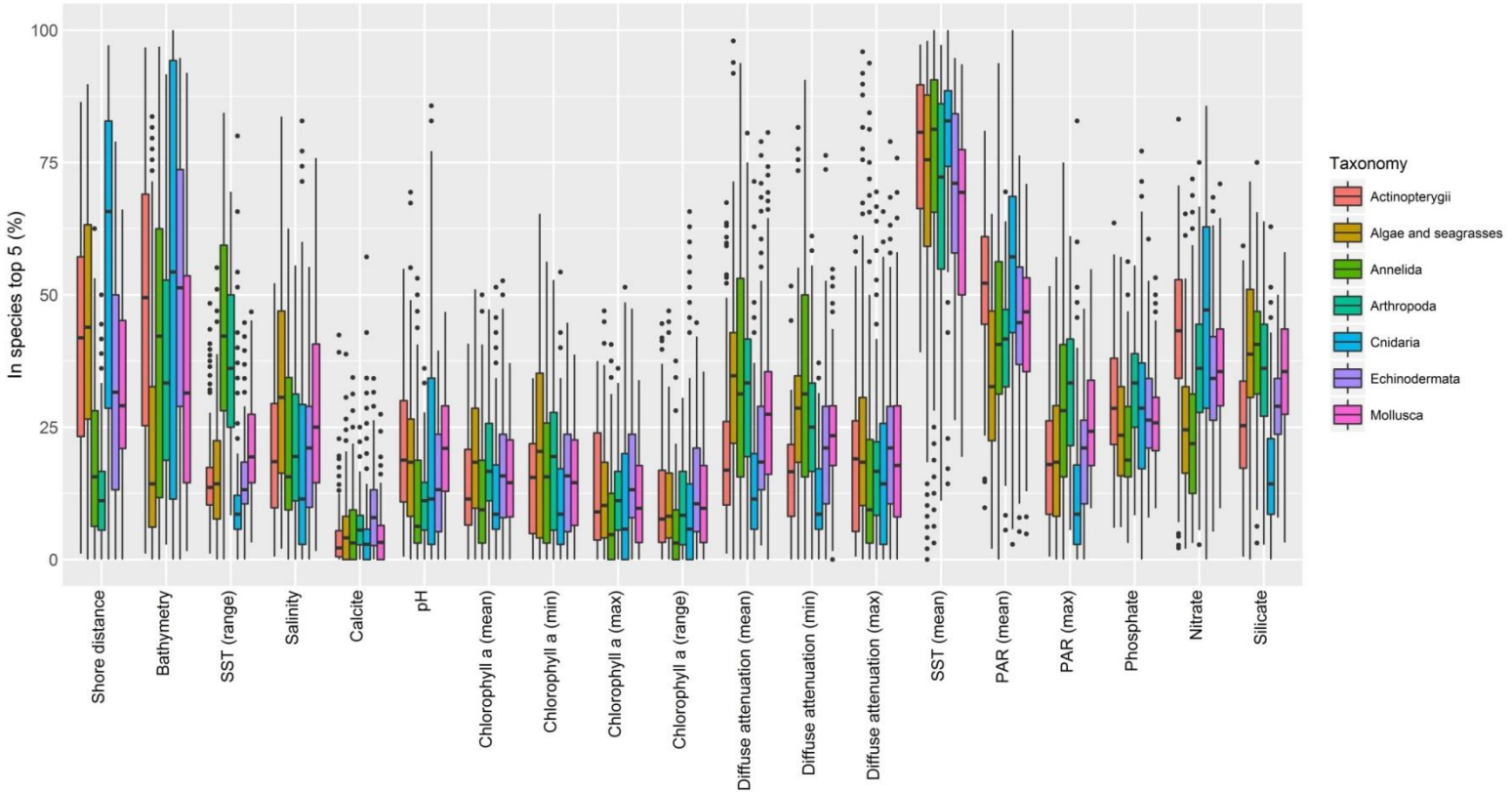


Figure S11. Percentage of species a predictor has a top 5 ranking in the different model setups for a selection of common taxonomic groups: Actinopterygii (red), algae and seagrasses (brown), Annelida (green), Arthropoda (cyan), Cnidaria (blue), Echinodermata (purple) and Mollusca (pink).

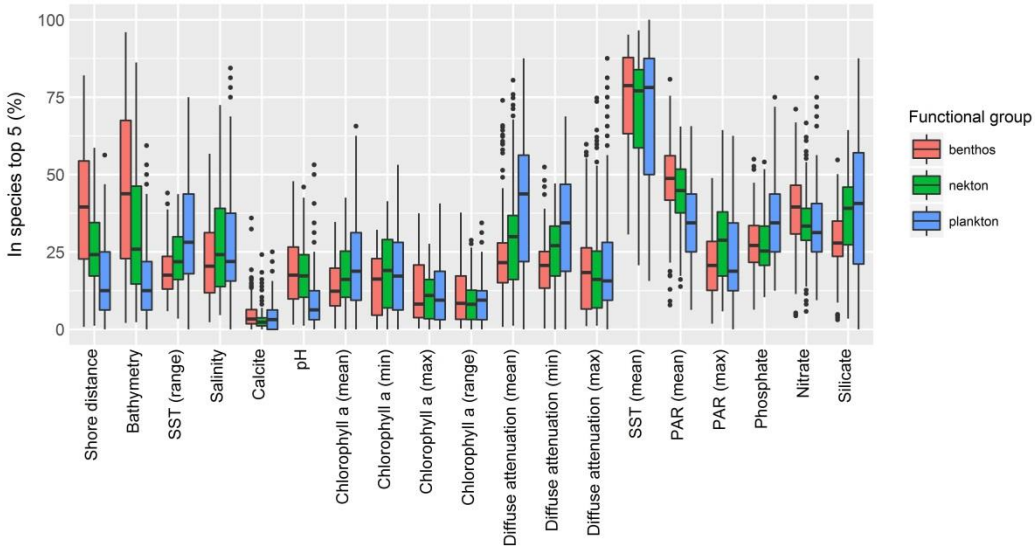


Figure S12. Percentage of species a predictor has a top 5 ranking in the different model setups for the different functional groups: benthos (red), nekton (green) and plankton (blue).

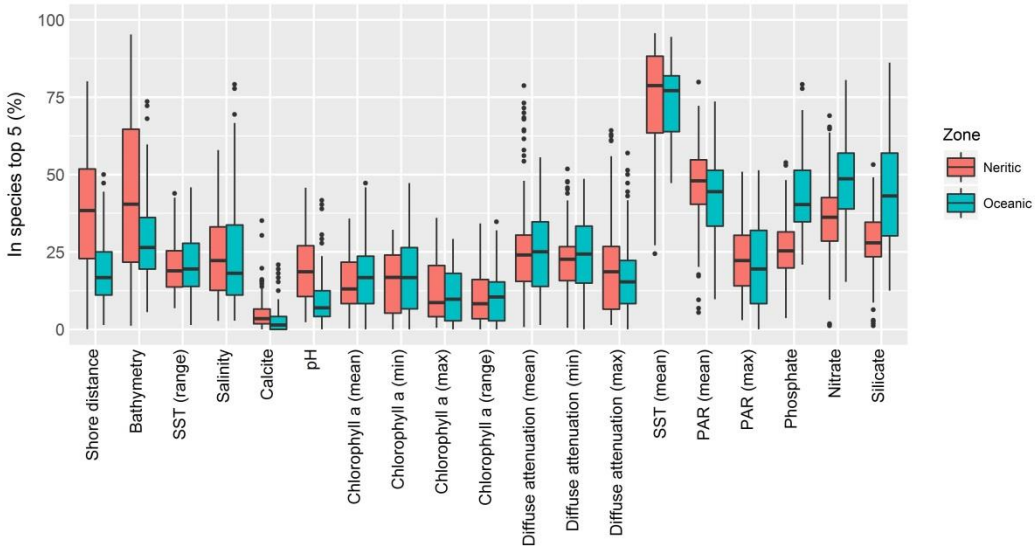


Figure S13. Percentage of species a predictor has a top 5 ranking in the different model setups for the different zones: neritic (red) and oceanic (blue).

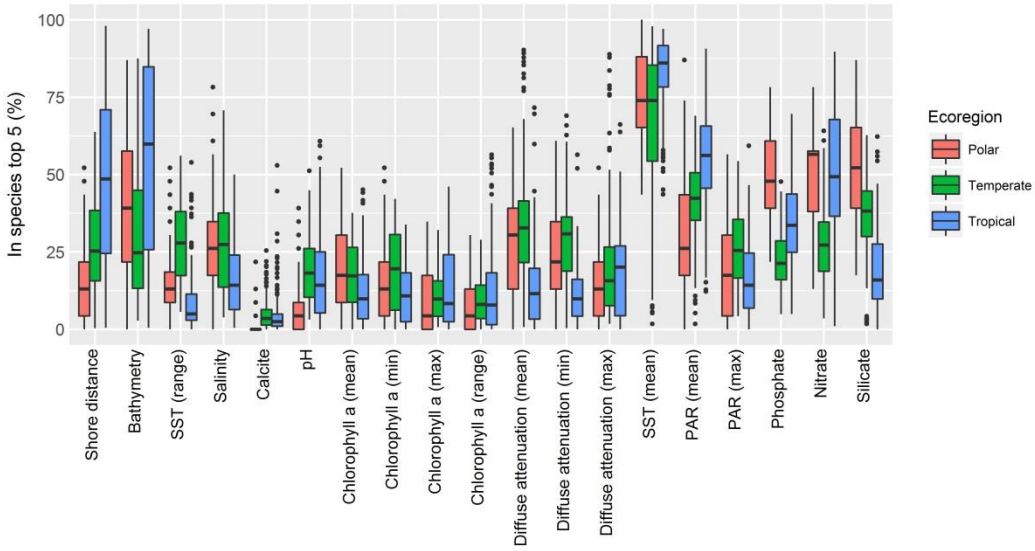


Figure S14. Percentage of species a predictor has a top 5 ranking in the different model setups for the different ecoregions: polar (red), temperate (green) and tropical (blue).

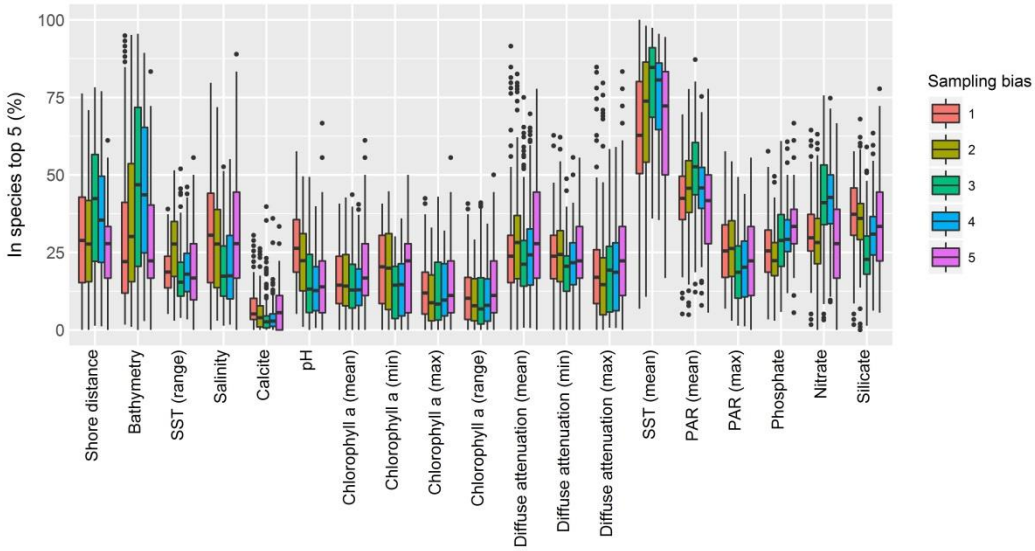


Figure S15. Percentage of species a predictor has a top 5 ranking in the different model setups for the different levels of sampling bias: 1 (low bias, red), 2 (brown), 3 (green), 4 (blue) and 5 (high bias, purple).

Chapter 5

Spatio-temporal patterns of introduced seaweeds in European waters, a critical review.

Samuel Bosch[1,2], Olivier De Clerck[1] and Frédéric Mineur[3]

[1]*Research Group Phycology, Biology Department, Ghent University, Krijgslaan 281/S8, 9000 Ghent, Belgium*
[2]*Flanders Marine Institute (VLIZ), InnovOcean site, Wandelaarskaai 7, 8400 Ostend, Belgium*
[3]*School of Biological Sciences, Queen's University, 97 Lisburn Road, Belfast BT9 7BL, United Kingdom*

Manuscript in preparation.

Abstract

Introductions of non-native species are a serious source of concern. A study on human-mediated transport of species demonstrated that rates of introductions have increased globally over the past centuries and averaged out over taxonomic groups there is no sign of saturation. Using seaweeds as a model group for the marine environment, a quantitative assessment of the temporal dynamics of primary and secondary introductions of seaweeds show that the rate of nonindigenous species being reported for the first time in European waters started declining since the beginning of the 1990's. To investigate whether this trend reflects a decline in the number of species being introduced or whether the discovery rate has declined because of factors other than the introduction rate, we analyzed trends in the literature of introduced seaweed species. Contrary to the rate of newly introduced species, the rate of the total number of records remained constant since 1990, with 115-120 records being recorded annually. The number of papers and authors increased spectacularly from 1970 to 2000 but shows a decrease from then onward. The combination of trends is interpreted as a decline in the rate new are yearly being introduced. Classifying introduced species according to geographical origin, the decline is mainly attributable to lower numbers of nonindigenous species with a NW Pacific origin being recorded from Europe, while the discovery rates of Lessepsian species or species native to Australasia has remained constant over the years. Given that livestock transfer of shellfish is the principal vector for the introduction of NW Pacific species, it appears that the increased awareness of authorities and stakeholders, and the implementation of policies dedicated to the prevention of introductions, reduce, but not prevent, the introduction of nonindigenous species.

Introduction

Over the last centuries thousands of species have dispersed outward from their native regions through human-mediated transport and have established populations in distant parts of the globe (Williams & Smith, 2007; Molnar et al., 2008). Many of these organisms have profoundly affected the abundance and diversity of native biota in the regions they have invaded (Vilà et al., 2010; Gallardo et al., 2016a), and in some cases they have had substantial economic impacts (Lovell et al., 2006; Holmes et al., 2009). A global analysis of temporal dynamics of species introductions by Seebens et al. (2017) demonstrated that the rate of introductions has significantly increased over the past centuries. Furthermore, over all taxonomic groups there are no signs of saturation and for most taxa the rate of introductions is still increasing. This trend has been linked to intensified global trade and transport (Seebens et al., 2013).

Because of differences in relative importance of the vectors for different groups of organisms in marine and terrestrial ecosystems, examining spatial and temporal patterns of specific taxonomic groups can inform policy makers about the effectiveness of targeted measures taken to mitigate the influx of nonindigenous species.

Here we present data on the patterns of marine species introductions and their spread at a European scale, using seaweeds as a model group for this assessment on marine environment. Representing one of the largest groups of marine aliens, constituting between 20 and 29% of all marine introduced species in Europe (Schaffelke et al., 2006), seaweeds are particularly fit for the purpose. Furthermore, they are attached and non-motile, reducing sampling errors caused by individual movement of target organisms. In addition, they comprise representatives of three major phyla, which, though widely divergent phylogenetically, have a series of convergent functional forms. In coastal systems, particularly on rocky shores, seaweeds are the dominant primary producers, playing a central structural and functional role in several habitats ranging from turfs to kelp forests (Mineur et al. 2015). Large-scale substitution of dominant native seaweeds with alien species may alter coastal productivity and food web structure, and therefore impact ecosystem services. Impact studies on invasive seaweeds have been carried out worldwide, and these have detected a range of ecological effects, mostly highlighting reduction in abundance of native biota (Williams & Smith, 2007). Maritime traffic and livestock transfer in aquaculture, in particular oyster cultivation, are usually regarded as the main vectors for primary introductions of alien seaweeds to Europe (Wallentinus, 2002; Mineur et al., 2012, 2015). A fragmented and to some extent incoherent policy

set up at EU and national levels, should prevent or limit introductions of marine organisms in European waters (Mineur et al. 2014). Directly relevant, although originally implemented only to mitigate the spread of diseases, are regulations of shellfish transfer within Europe and restrictions on the import of livestock from outside of Europe (Mineur et al., 2014). Although posing a reduced risk toward the introduction of nonindigenous seaweed species, similar measures have been taken to reduce the risk of introductions by hull fouling or ballast water discharge (Flagella et al., 2007; Mineur et al., 2008). Despite the environmental risk imposed by nonindigenous seaweeds, a comprehensive overview of the spatial and temporal dynamics of introductions in Europe is lacking. Overall scarcity of baseline data, which species have been introduced, the rates of introductions versus the rates of discovery and regional patterns of introduced species, fall short to test the effectiveness of prevention policies and therefore limit prevention of further introductions.

To address this knowledge-gap we compiled a database of nonindigenous seaweed species and distribution records in Europe, their likely origin and introduction vectors. These data are used to provide a quantitative assessment of the spatio-temporal dynamics of primary and secondary introductions in Europe. A comparison of discovery rates with statistics of the number of papers and the size of the phycological community that reports on nonindigenous seaweeds is used to infer conclusions on the introduction rates.

Materials and methods

Data compilation

We compiled a database of non-native marine seaweed species records reported from the Northeast Atlantic, Mediterranean and Macaronesian coasts (the Azores, Canary Islands, Cape Verdes, and Madeira). We report the year of the first report of the nonindigenous species in these three regions. Where possible this date refers to the year the species was first observed. In the absence of such information the date refers to the year the first record was published. If unclear when the species was introduced a question mark is added. The dataset, which includes published and unpublished records produced by various local and European research projects, builds on previous lists by Mineur et al. (2010) and Verlaque et al. (2015). Data from Macaronesia are based on Gil-Rodríguez et al. (2003) and Gallardo et al. (2016).

Refinements to previous lists were needed because in the past the term introduced species has likely been used too liberally. The introduced nature of certain species

was sometimes based on scanty evidence. In the present database we hope to remedy this by critically revising the list of nonindigenous seaweeds and by explicitly expressing confidence in the taxonomy and introduced nature of the species. First, species are considered nonindigenous when their presence in a given region is the result of a displacement linked to human activities either through a transport vector, or through the removal of a physical barrier, e.g. between the Red Sea and the Mediterranean Sea through the opening of the Suez Canal. The dataset also includes indigenous European species that have demonstrably become displaced within Europe as a result of human-mediated exchanges. Examples include exchanges of species between Atlantic and Mediterranean shores. However, true cosmopolitan species, whose current distribution may have been shaped by human transport, were omitted. Second, given widespread taxonomic uncertainty that surrounds many algal names we assigned an index of taxonomic accuracy for every species. We assigned a 'high' score to accepted nominal species that were not shown to be a species complex based on molecular studies in their European introduced or native ranges. A high score was also assigned to species for which, so far, there is sufficient confidence in unambiguous identification based on morphology. Conversely, species that belong to an understudied complex of cryptic species are assigned a low score. Related to, but not necessarily equivalent to taxonomic uncertainty, is the confidence that a species is indeed nonindigenous in European waters. To this end, we introduced a separate category, 'xenoticity'. In addition, we indicate the introduction status on an ordinal scale, ranging from not recorded, to recorded but not known to be established, likely established with recurrent observations or abundant in restricted areas, to widespread and abundant. If doubt exists regarding the introduced nature of a species in any of the three regions, this is indicated as 'potentially native'. Third, for each species an estimate is provided for their native biogeographical range. To do this, global distribution data were obtained from Algaebase (Guiry & Guiry, 2017). Indices of confidence of native ranges were assigned to each marine biogeographic region. Null values correspond to absence of records, medium indices to the presence of the species without a high confidence in stating if this biogeographic realm is the native range, while high indices are given to biogeographic regions that can be unambiguously determined as the origin, in the native range, of the populations present in Europe. We note that European populations may however have transited through other marine realms by secondary introductions.

Statistical analyses

Distribution data were gridded on raster cells of 100 km x 100 km (10,000 km²). The statistical analysis of spatio-temporal patterns was restricted to records until the year

2010, to avoid a potential bias due to lags in the reporting of nonindigenous seaweed species. We fitted three functions (linear, power and logistic) to the cumulative plots of the number of introduced species and the number of distribution records with the R package *minpack.lm*. The distribution of the AIC values and the midpoint of the logistic curve have been estimated from 1000 bootstrap samples. Additionally we fitted a local regression (LOESS) to these same cumulative plots and calculated the yearly rate of change in the number of introduced species and distribution. The span was calculated automatically by minimizing the AICc of the LOESS curve using the R package *fANCOVA*. Visualisation of hotspots of introductions is based on binned kernel density maps for the first record of every introduced species and for all distribution records with the R package *KernSmooth*.

Results

List introduced seaweeds: uncertainty in the numbers

In total 153 seaweed species have been listed as introduced in Europe, of which 104 species are red algae (Rhodophyta), 29 brown algal species (Phaeophyceae) and 20 belonging to the green lineage (Chlorophyta, Charophyta) (Fig. 1A; Table 1). However, an unequivocal link between specimens found in Europe with specimens in the native range has only been established for about half of these species. Given the widespread nature of cryptic and pseudocryptic diversity in algae in general it should come as no surprise that molecular studies have been substantially revising our view on many introduced species. For example, several species have been described from Europe which later turned out to represent introduced species. For example, *Dictyota cyanoloma* was described as a new species from the Mediterranean Sea and Macaronesia as recently as 2010 (Tronholm et al., 2010), but subsequent collecting efforts in Australia revealed that the species actually represent a cryptic introduction (Aragay et al., 2016; Steen et al., 2017). Similarly, *Porphyra olivii* described by Brodie et al. (2007) from the Mediterranean belongs to the same species as *Pyropia koreana* (Vergés et al., 2013). Obviously determining the nonindigenous nature of a species becomes much more difficult if introductions took place long ago as is the case for *Codium fragile* subsp. *fragile* and *Neosiphonia harveyi* which were established in Europe already by the mid-19th century as evidenced by herbarium records (McIvor et al., 2001; Provan et al., 2008). A puzzling case is formed by several taxa with clear Indo-Pacific affinities which appeared in the Mediterranean Sea prior to the opening of the Suez Canal in 1869, *Acanthophora nayadiformis, Asparagopsis taxiformis* and *Ganonema farinosum.* At least for *A. taxiformis*, such a counterintuitive temporal pattern can be explained by the presence of two cryptic lineages, which include a

native strain present in the Mediterranean Sea prior to the opening of the Suez canal and a more recent introduction of an invasive strain (Chualáin et al., 2004; Andreakis et al., 2007). Such cases highlight the difficulty in establishing whether a species is introduced in Europe. Overall, only for about half of the species (54%) listed in Table 1, there is strong evidence that they are nonindigenous in Europe. For the remaining half the evidence is mediocre (35%) to weak (11%) at least. It should be noted that taxonomic uncertainty does not per se correlate with xenoticity. For several species there is good evidence that they are indeed nonindigenous, however, the taxonomy of the group is still not sufficiently established to be certain regarding the correct name of the species. Taxonomic uncertainty is not necessarily restricted to diminutive species which have been observed sporadically. *Agardhiella subulata* is a good example, its nonindigenous nature is not questioned, however, according to some authors the species should be identified as *Sarcodiotheca gaudichaudii* (Montagne) P.W.Gabrielson (Stegenga & Karremans, 2015). Likewise the correct taxonomic status of many of the introduced *Caulerpa* species found in the Eastern Mediterranean Sea (e.g. *C. lamourouxii*, *C. mexicana*, *C. scalpelliformis*) needs further study (Verlaque et al., 2000, 2015; Belton et al., 2014).

Taxonomic uncertainty and uncertainty regarding the introduced nature of seaweed species is most prevalent in the Macaronesian region. Out of 57 species present in Macaronesia no less than 27 have been given a low taxonomic accuracy score. Although several factors likely contribute to this uncertainty, the geographical location of the region, bordering the tropical Atlantic, contributes significantly to the difficulty in interpretation of the nonindigenous nature of species. Many tropical and subtropical taxa are reported from all major ocean basins. Very often these taxa represent (pseudo-)cryptic species complexes with the individual species being either range-restricted or widespread themselves. The lack of accurate baseline data regarding species boundaries and distributions makes it particularly hard to distinguish native from introduced seaweeds. Examples include *Caulerpa* spp., *Hypnea* spp., *Galaxaura rugosa*, *Ganonema farinosum*.

Table 1. Overview of the nonindigenous seaweeds in Europa, with indication of their presence in the NE Atlantic, Mediterranean Sea and Macaronesia (numbers = year of the first record, NA = not recorded, NT = Native, ? = uncertain | color codes: green = recorded but not known to be established, orange = likely established, recurrent observation to abundant in restricted areas, red = widespread and abundant, invasive, blue = (potentially) native). Taxonomic uncertainty is indicated in white = low, green = high. Xenoticity expresses the certainty that the species is indeed introduced (white = low, pale green = medium, dark green = high). Displacement (R = from remote geographical area; L = Erythrean migrant; M = NE Atlantic to Mediterranean , A = range extension in the NE Atlantic; U = unknown or ambiguous). Origin denotes the most likely native area (white = unlikely, pale green = potential, dark green = high likelihood).

		Region			Uncertainty			Origin					
Taxonomic name	Nr of Records	NE Atlantic	Mediterranean	Macaronesia	Taxonomic unc.	Xenoticity	Displacement	NW Pacific	Lessepsian	Australasia	W Atlantic	NE Atlantic	Indo-Pacific
Phaeophyta	**1762**												
Acrothrix gracilis	1	NT	1998	NA	0	0	R	1	1	0	1	1	0
Ascophyllum nodosum	2	NT	2009	NA	1	2	R	0	0	0	0	2	0
Botrytella parva	2	NA	1996	NA	0	0	R	1	0	0	0	1	0
Chorda filum	3	NT	1981	NA	1	2	M	0	0	0	0	2	0
Cladosiphon zosterae	1	NT	1998	NT	1	1	M	0	0	0	0	1	0
Colpomenia peregrine	172	1905	1918	1965	1	2	R	2	0	0	0	0	0
Corynophlaea verruculiformis	6	1994	NA	NA	0	1	R	2	0	0	0	0	0
Corynophlaea cystophorae	0	NA	NA	1993	0	1	R	2	0	0	0	0	0
Desmarestia viridis	6	NT	1947	NA	1	1	U	1	0	0	0	1	0
Dictyota cyanoloma	286	2008	1935	2007	1	2	R	0	0	2	0	0	0
Ectocarpus siliculosus var. hiemalis	1	NA	1998	NA	0	1	M	0	0	0	0	2	0
Fucus evanescens	33	1883	NA	NA	1	2	U	1	0	0	2	0	0
Fucus serratus [Iceland and Faroes]	58	1897	NA	NA	1	2	A	0	0	0	0	2	0
Fucus spiralis	1	NT	1987	NA	1	2	M	0	0	0	0	2	0
Halothrix lumbricalis	4	NT	1978	NA	0	1	U	1	0	0	0	1	0
Leathesia marina	3	NT	1905	NA	1	2	M	0	0	0	0	2	0
Padina boergesenii	20	NA	1965	NA	1	1	L	0	1	1	1	0	1
Padina boryana	1	NA	1993	NA	1	1	L	0	1	1	0	0	1
Petalonia binghamiae	20	NA	NA	1980	0	1	U	1	0	1	1	0	1
Punctaria tenuissima	6	NT	1957	NA	0	2	U	0	0	0	1	2	0
Pylaiella littoralis	1	NT	1960	NA	1	2	M	0	0	0	0	2	0
Rugulopteryx okamurae	1	NA	2002	NA	1	2	R	2	0	0	0	0	0
Saccharina japonica	2	NA	1976	NA	1	2	R	2	0	0	0	0	0
Sargassum muticum	924	1972	1981	NA	1	2	R	2	0	0	0	0	0
Scytosiphon dotyi	12	1987	1977	1993	0	1	R	1	0	0	1	0	0
Spatoglossum variabile	2	NA	1944	NA	0	1	L	0	2	0	0	0	1
Sphaerotrichia firma	1	NA	1970	NA	1	2	R	2	0	0	0	0	0
Stypopodium schimperi	22	NA	1973	1997	1	2	L	0	2	0	0	0	0
Undaria pinnatifida	171	1982	1971	NA	1	2	R	2	0	0	0	0	0
Chlorophyta	**721**												
Caulerpa chemnitzia	14	NA	1926	NT?	0	1	L	0	2	0	0	0	1
Caulerpa cylindracea	106	NA	1990	2002	1	2	R	0	0	2	0	0	0
Caulerpa lamourouxii	12	NA	1951	NA	1	2	L	0	2	0	0	0	1
Caulerpa mexicana	14	NA	1941	NT?	0	2	L	0	2	0	0	0	1
Caulerpa scalpelliformis	14	NA	1929	NA	0	2	L	0	2	0	0	0	1
Caulerpa taxifolia	87	NA	1984	NA	1	2	R	0	0	2	0	0	0
Caulerpa taxifolia var. *distichophylla*	17	NA	2006	NA	1	2	R	0	0	2	0	0	0
Cladophora herpestica	10	NA	1948	NA	0	0	L	1	2	1	0	0	1

		Region			Uncertainty			Origin					
Taxonomic name	Nr of Records	NE Atlantic	Mediterranean	Macaronesia	Taxonomic unc.	Xenoticity	Displacement	NW Pacific	Lessepsian	Australasia	W Atlantic	NE Atlantic	Indo-Pacific
Cladophora patentiramea	1	NA	1991	NA	0	1	L	0	1	1	0	0	1
Codium arabicum	5	2003	2007	NA	1	2	L	1	2	0	0	0	1
Codium fragile subsp. *fragile*	397	1845	1950	1990	1	2	R	2	0	0	0	0	0
Codium parvulum	2	NA	2004	NA	1	2	L	0	2	0	0	0	1
Codium taylorii	23	2004	1955	?	0	1	R	0	1	0	1	0	1
Derbesia boergesenii	1	NA	1972	NA	0	1	L	0	2	0	0	0	1
Derbesia rhizophora	2	NA	1984	NA	0	2	R	2	0	0	0	0	0
Halimeda incrassate	3	NA	2011	2005	1	2	R	0	0	0	2	0	0
Neomeris annulata	1	NA	2003	NA	1	2	L	1	2	1	1	0	1
Ulva pertusa / U. australis	10	1993	1984	?	1	1	R	2	0	1	1	1	0
Ulvaria obscura (Kützing)	2	NT	1985	NA	0	1	U	1	0	0	0	1	0
Charophyta	**17**												
Chara connivens	17	1979	NA	1975	0	1	A	0	0	0	0	0	0
Rhodophyta	**2340**												
Acanthophora nayadiformis	56	NA	1808	NA	1	1	L	0	2	0	0	0	1
Acrochaetium balticum	1	1998	NA	NA	0	0	A	0	0	0	0	2	0
Acrochaetium robustum	1	NA	1944	NA	0	0	L	0	2	1	0	0	0
Acrochaetium spathoglossi	3	NA	1944	NA	0	0	L	0	2	1	0	0	0
Acrochaetium subseriatum	3	NA	1944	NA	0	0	L	0	2	1	0	0	0
Acrothamnion preissii	62	NA	1969	NA	1	2	R	1	0	2	0	0	0
Agardhiella subulata	1	1973	1984	NA	0	1	R	1	0	0	1	0	0
Aglaothamnion feldmanniae	3	NT	1975	NA	0	1	M	0	0	0	0	1	0
Aglaothamnion halliae	24	1960	NA	NA	0	1	R	0	0	0	1	0	0
Ahnfeltiopsis flabelliformis	3	NA	1994	NA	0	2	R	2	0	0	0	0	0
Anotrichium furcellatum	39	1922	1939	1930	0	0	U	1	0	0	0	0	0
Antithamnion amphigeneum	38	1995	1989	NA	1	2	R	0	0	2	0	0	0
Antithamnion densum	21	1992	NA	1990	0	0	R?	1	0	0	1	1	0
Antithamnion diminuatum	2	NA	NA	1988	1	0	R	0	0	2	0	0	0
Antithamnion nipponicum / A. hubbsii	10	2003	1988	NA	0	1	R	2	0	1	0	0	0
Antithamnionella boergesenii	14	2004	1937	1921	1	1	R	0	0	0	2	1	0
Antithamnionella elegans	85	1961	1882	NA	0	1	R	2	0	0	0	0	0
Antithamnionella spirographidis	79	1927	1911	1974	0	0	R	2	0	0	0	0	0
Antithamnionella sublittoralis	5	NA	1980	NA	0	1	R	1	0	0	0	0	0
Antithamnionella ternifolia	113	1906	1926	NA	0	1	R	0	0	2	0	0	0
Apoglossum gregarium	11	NA	1992	NA	1	2	R	1	0	0	1	0	0
Asparagopsis armata	386	1923	1923	1965	1	2	R	0	0	2	0	0	0
Asparagopsis taxiformis [invasive strain]	39	2004	?	NT?	1	2	R	0	1	2	0	0	1
Bonnemaisonia hamifera	281	1893	1909	1930	1	2	R	2	0	0	0	0	0
Botryocladia madagascariensis	19	NA	1991	1988	0	2	R	0	0	0	0	0	1
Caulacanthus okamurae	23	1986	2004	NA	1	2	R	2	0	0	0	0	0
Ceramium bisporum	4	NA	2001	NA	0	1	R	0	0	0	2	0	0
Ceramium strobiliforme	15	NA	1991	1992	0	0	R	0	0	0	0	0	0
Chondracanthus chamissoi	1	2009	NA	NA	1	2	R	2	0	0	0	0	0
Chondria curvilineata	5	NA	1981	NA	0	1	R	0	0	0	2	0	1
Chondria polyrhiza	2	NA	1982	NA	0	0	R	0	0	0	2	0	1
Chondria pygmaea	14	NA	1974	NA	0	2	R	0	2	0	0	0	1
Chondrus giganteus	2	NA	1994	NA	1	2	R	2	0	0	0	0	0
Chrysymenia wrightii	15	2005	1978	NA	1	2	R	2	0	0	0	0	0
Colaconema codicola	7	1957	1952	NT	0	1	U	0	0	0	1	1	0
Colaconema dasyae	2	1983	NA	NA	1	1	R	0	0	0	0	0	0

		Region			Uncertainty			Origin					
Taxonomic name	Nr of Records	NE Atlantic	Mediterranean	Macaronesia	Taxonomic unc.	Xenoticity	Displacement	NW Pacific	Lessepsian	Australasia	W Atlantic	NE Atlantic	Indo-Pacific
Cryptonemia hibernica	30	1971	NA	NA	1	1	R	0	0	0	0	1	0
Dasya baillouviana	21	1950	NT	NT	0	1	U	0	0	1	1	1	1
Dasya sessilis	43	1989	1984	NA	1	2	R	2	0	0	0	0	0
Dasysiphonia japonica	61	1994	1998	NA	1	2	R	2	0	0	0	0	0
Devaleraea ramentacea	1	1975	NA	NA	1	2	A	0	0	0	0	1	0
Ezo epiyessoense	1	1983	NA	NA	1	1	R	2	0	0	0	0	0
Fredericqia deveauniensis	4	1850	NA	NA	1	2	R	0	0	0	2	0	0
Galaxaura rugosa	3	NA	1990	NT	1	2	L	0	2	1	1	0	1
Ganonema farinosum	10	NA	1808	NT	0	0	L	0	2	1	1	0	1
Gelidium vagum	4	2010	NA	NA	1	2	R	2	0	0	0	0	0
Goniotrichopsis sublittoralis	11	1975	1989	NA	0	1	R	0	0	0	0	0	0
Gracilaria arcuata	9	NA	1931	NA	0	1	L	1	2	1	0	0	1
Gracilaria disticha	2	NA	1924	NA	0	2	L	0	2	0	0	0	1
Gracilaria vermiculophylla	80	1997	2008	NA	1	2	R	2	0	0	0	0	0
Gracilariopsis chorda	1	2010	NA	NA	1	2	R	2	0	0	0	0	0
Grateloupia asiatica	11	2010	1984	NA	1	2	R	2	0	0	0	0	0
Grateloupia imbricate	5	2014	NA	2006	1	2	R	2	0	0	0	0	0
Grateloupia patens	3	NA	1994	NA	1	2	R	2	0	0	0	0	0
Grateloupia subpectinata	20	1947	1990	1983	1	2	R	2	0	0	0	0	0
Grateloupia turuturu	85	1969	1982	1983	1	2	R	2	0	0	0	0	0
Griffithsia corallinoides	9	NT	1964	NA	0	1	U	1	0	0	0	2	0
Gymnophycus hapsiphorus	7	NA	NA	1989	0	1	R	0	0	1	0	0	0
Herposiphonia parca	2	2005	1991	NA	0	1	R	2	0	0	1	0	1
Hypnea anastomosans	2	NA	2008	NA	1	1	L	0	2	0	0	0	1
Hypnea cornuta	6	NA	1896	NA	1	2	?	1	2	0	0	0	1
Hypnea flagelliformis	1	NA	1956	?	0	2	U	1	2	0	0	0	1
Hypnea flexicaulis	3	NA	2009	NT	1	2	L	2	0	0	0	0	0
Hypnea musciformis	11	2003	NT	NT	0	0	U	0	1	1	1	0	1
Hypnea spinella	20	NA	1926	NT	0	1	?	1	1	1	1	0	1
Hypnea valentiae	3	NA	1996	NT	0	2	R	1	1	0	0	0	1
Laurencia brongniartii	1	1989	NA	NT	0	2	R	1	0	1	0	0	1
Laurencia caduciramulosa	11	NA	1991	NT	0	1	R	1	0	0	1	0	1
Laurencia okamurae	2	NA	1984	NA	1	2	R	2	0	0	0	0	0
Lithophyllum yessoense	1	NA	1994	NA	1	2	R	2	0	0	0	0	0
Lomentaria hakodatensis	26	1984	1978	NA	1	2	R	2	0	0	0	0	0
Lophocladia lallemandii	52	NA	1908	NA	1	2	L	1	2	1	0	0	1
Mastocarpus stellatus [Helgoland]	1	1983	NA	NT	1	1	A	0	0	0	0	2	0
Monosporus indicus	5	NA	2015	NA	0	1	L	0	2	0	0	0	0
Neosiphonia harveyi	109	1832	1958	1990	1	2	R	2	0	0	0	0	0
Nitophyllum stellatocorticatum	2	NA	1984	NA	1	2	R	2	0	0	0	0	0
Pachymeniopsis gargiuli	6	NA	2000	2007	1	2	R	2	0	0	0	0	0
Pachymeniopsis lanceolata	11	NA	1982	NA	1	2	R	2	0	0	0	0	0
Palisada maris-rubri	2	NA	1990	NA	0	0	L	0	2	0	0	0	1
Pikea californica	22	1967	NA	NA	1	2	R	0	0	0	0	0	0
Plocamium secundatum	4	NA	1976	NA	0	0	U	0	0	0	0	0	0
Polyopes lancifolius	2	2008	NA	NA	1	1	R	0	0	0	0	0	0
Polysiphonia atlantica	1	NT	1972	NT	1	1	M	1	0	1	1	2	0
Polysiphonia morrowii	23	1993	1997	NA	1	2	R	2	0	0	0	0	0
Polysiphonia paniculata	10	NA	1967	NA	0	1	R	0	0	0	0	0	0
Polysiphonia schneideri	2	2010	NA	NA	1	2	R	0	0	0	2	0	0
Predaea huismanii	1	NA	NA	1990	0	1	U	0	0	1	0	0	1
Pterosiphonia pinnulata	2	1990	NT	NA	0	2	U	1	0	0	0	0	0

		Region			Uncertainty			Origin					
Taxonomic name	Nr of Records	NE Atlantic	Mediterranean	Macaronesia	Taxonomic unc.	Xenoticity	Displacement	NW Pacific	Lessepsian	Australasia	W Atlantic	NE Atlantic	Indo-Pacific
Pterosiphonia tanakae	2	2005	1998	NA	1	2	R	2	0	0	0	0	0
Pyropia koreana	6	NA	2007	NA	1	2	R	2	0	0	0	0	0
Pyropia suborbiculata	12	2010	2010	1993	1	2	R	2	0	0	0	0	0
Pyropia yezoensis	2	1984	1976	NA	0	2	R	2	0	0	0	0	0
Rhodophysema georgei	1	NT	1978	NA	0	1	M	1	0	0	0	2	0
Rhodymenia erythraea	1	NA	1948	NA	0	2	L	0	2	0	0	0	0
Sarconema filiforme	11	NA	1945	NA	0	2	L	0	2	1	0	0	1
Sarconema scinaioides	2	NA	1945	NA	0	2	R	0	2	1	0	0	1
Scageliopsis patens	12	2004	NA	1989	0	2	R	0	0	2	0	0	0
Solieria dura	3	NA	1944	NA	0	1	L	0	2	0	0	0	1
Solieria filiformis	4	1980	1988	2002	0	1	R	0	1	1	1	0	1
Solieria sp. [non described]	5	2005	2011	NA	0	2	R	1	0	1	1	0	0
Spongoclonium caribaeum	21	1973	1974	1980	0	1	U	1	1	1	1	0	1
Symphyocladia marchantioides	10	2004	1984	1971	0	1	R	1	0	1	0	0	0
Vertebrata fucoides	2	NT	1988	NT	1	1	M	0	0	0	1	2	0
Womersleyella setacea	92	NA	1986	1983	0	2	R	0	0	2	0	0	1

Records – temporal trends

A total of 4900 distribution records from published and non-published sources were compiled for this study. Nearly half of the distribution records (47%) concern the five most represented species: *Sargassum muticum* (925 records), *Codium fragile* subsp. *fragile* (397 records), *Asparagopsis armata* (386 records), *Bonnemaisonia hamifera* (279 records) and *Dictyota cyanoloma* (286 records) (Fig.1). These species have been the focus of dedicated research projects, are usually large in size, easy to identify on the field, and often have considerable population sizes. At the other end of the spectrum, there are 73 species with less than 5 distribution records of which 27 species have only been recorded once.

Breaking down the number of introduced species into the Mediterranean Sea, Atlantic shores and Macaronesia reveals that 63% of the introduced species have been reported for the first time in the Mediterranean Sea, 27% in the NE Atlantic and 10% in Macaronesia (Fig. 2A). Twelve introduced species are shared among the three regions, while one third of the species occurs nowadays in 2 regions (Fig. 2B). Most species are shared between the Mediterranean and the NE Atlantic (30 species), while a surprisingly low 5 species are shared between the Mediterranean and Macaronesia. The ratio of species belonging to red, green and brown lineages is approximately the same for the three regions but the Atlantic area has less introduced green and brown species than the Mediterranean area.

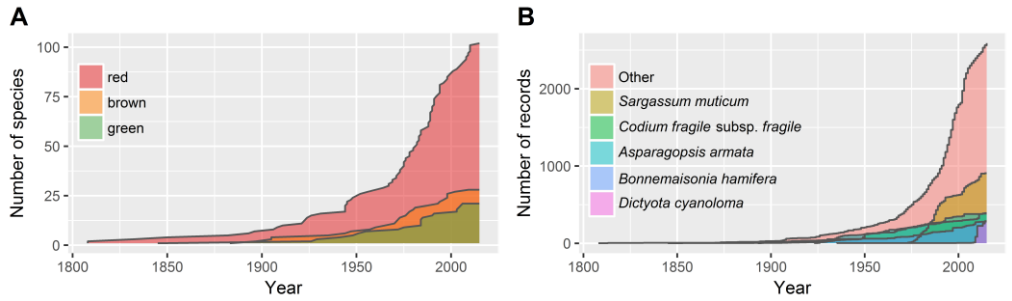


Figure 1. Number of introduced species and records through time in the whole study area. The left plot (A) shows the number of seaweeds introduced since 1800 for the red, brown and green classes. The right plot (B) shows the number of distribution records since 1800 of particularly well-studied introduced species in Europe: *Sargassum muticum* , *Codium fragile* subsp. *fragile* , *Asparagopsis armata* , *Bonnemaisonia hamifera* , *Dictyota cyanoloma* . In pink are the number of records for the remaining species. All curves are cumulative and superimposed.

A.

Region	Total	First	Red	Green	Brown
Mediterranean Sea	121	97	77	20	24
NE Atlantic	66	41	53	5	8
Macaronesia	31	15	21	4	6

B.

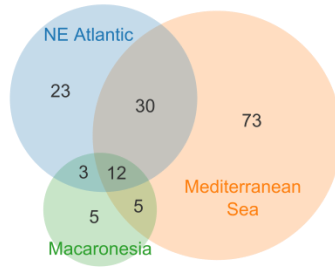


Figure 2. The left table (A) shows the total number of species reported in each European area; the number of species that has been reported for the first time in each European area, and the breakdown by red, green and brown algae of the total number of introduced species. The figure on the right (B) shows a Venn diagram of the number of the introduced species in the different areas.

The number of introduced species from 1950 to 2010, as represented by the date of the first record in Europe, was best fitted with a logistic curve (Fig. 3A). Likewise the total number of records of introduced species were also best presented by a logistic curve (Fig. 3B). For the bootstrapped AIC values of the different fitted curves we refer to Fig. S1 in Supporting information. The logistic curve implies that the number of new introduced species which are discovered is declining. Likewise, the accumulation rate of the number of distribution records of introduced species is also declining, albeit that the trend is less pronounced compared to the first record curve. These observations are confirmed by the rate of introduced species (Fig. 3C) which peaked in 1991, the sampling rate (Fig. 3D) which peaked in 1997 and by the midpoints of the logistic curves: 1986 for the number of introduced species and 1996 for the number of distribution records (Fig. S2 in Supporting information). The decrease in accumulation rate of nonindiginous seaweed species in Europe is at odds with general

trends as reported by Seebens et al. (2017) who observed a continuous rise in first record rates since 1800 for all groups of organisms except mammals and fishes.

Because the rate of discoveries of species are influenced by factors other than introductions (Costello & Sollow 2003), we also quantified the seaweed sampling effort along European coasts. We used the number of papers and number of unique authors reporting introduced seaweed species as a proxy for sampling effort (Fig. 4). These graphs disprove the idea that a decline in collecting or reporting effort underlies the slowdown in the number of first records.

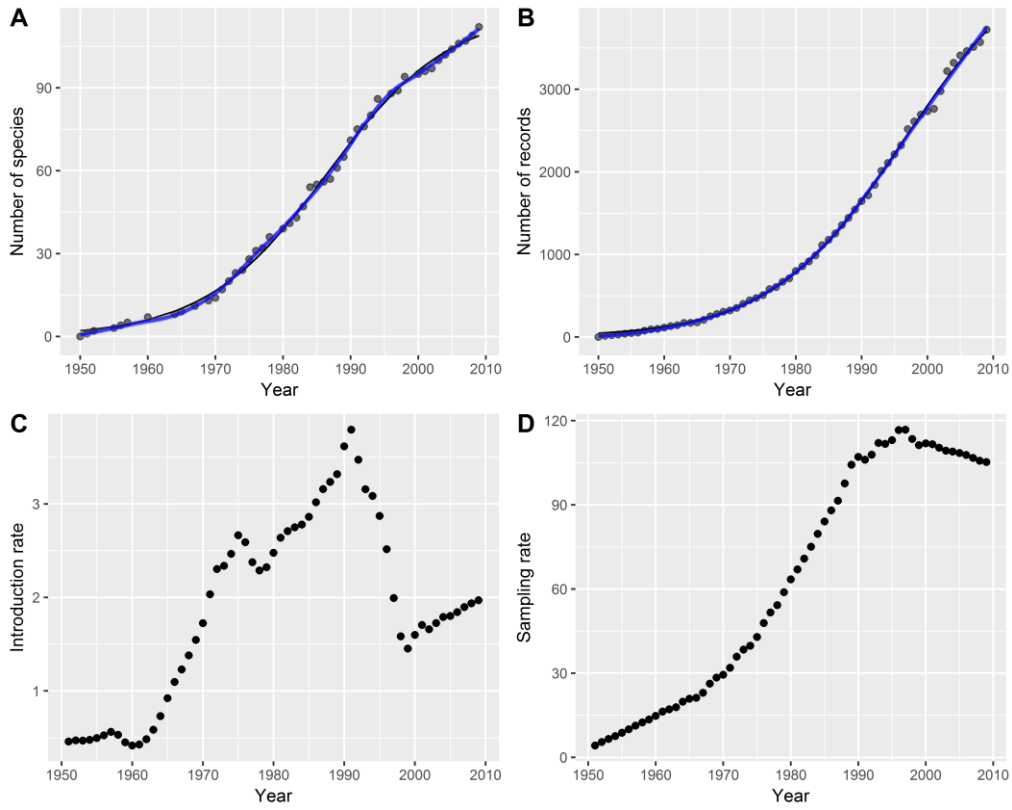


Figure 3. LOESS (blue) and logistic (black) curve fitting of the cumulative number of introduced species reported for the first time in Europe between 1950 and 2010 (A) and for the number of reported distribution records over all introduced species in the same area and period (B). The introduction rate (C) and the sampling rate (D) were calculated based on the fitted LOESS curves.

Figure 4. Number of unique authors per decade (A) and number of publications (B) reporting introduced species as a proxy for European sampling effort of introduced macroalgae.

The fact that such a decline in the introduction rate of non-indingenous species is apparently not shared with the majority of taxa or across geographic regions (Seebens et al. 2017), may reflect the somewhat atypical case of seaweed introductions and the success of measures aimed at mitigating new introductions. Contrary to most marine taxa, hull fouling and ballast water seem to play a relatively minor role only in the displacement of seaweeds across the globe. A disproportionate number of non-indigenous seaweed species appears to have been introduced through import of oyster stocks (Verlaque et al., 2007). In the late 1960s and early 1970s, disease outbreaks in Europe affecting oyster populations caused a major disruption of production. Mitigation procedures involved massive imports of oyster stock from the species' native range in the northwestern Pacific in the 1970s (Mineur et al., 2014). Alongside such stock imports non-native marine species were imported in great numbers from the northwestern Pacific to Europe. The accumulation curves of first records, which keep rising until the mid 1980s, mimick these imports. However it appears that European directives which authorizes all movements inside Europe and restrict shellfish stock imports from outside Europe reduce, but not prevent, the introduction of additional seaweed species.

Introduction hotspots

The importance of aquaculture toward introductions of seaweeds is reflected in the distributions of the first record of each species in Europe. A kernel density map (Fig. 5A) clearly shows the Thau lagoon, with 30 reports of first introductions in Europe (25%), as one of the major introduction hotspots in Europe. In total 58 species, 32% of the total seaweed diversity or 48-99% of the biomass, have been introduced in the Thau Lagoon (Boudouresque et al., 2010). The Thau lagoon is the epicentrum of oyster cultivation in the Mediterranean Sea. However, the oyster farmers rely entirely

on the import of juvenile oysters from other regions, European or non-European, because the lagoon is not suitable for oyster breeding. These continuous transfers result in astonishingly high numbers of introduced species. Upon closer examination, the Thau lagoon as well as other Mediterranean lagoons (Mar Piccolo, Venice lagoon), stand out with respect to introduction of native Atlantic species in the Mediterranean Sea (e.g. *Ascophyllum nodosum*, *Chorda filum, Cladosiphon zosterae, Pylaiella littoralis, Vertebrata fucoides*). A low native diversity due to the low occurrence of natural hard substrata in lagoons, and relatively recent construction of hard substrata for aquaculture purposes, concomitant with transfers of lifestock which seed the new substrata, makes these habitats hotspots for nonindigenous species (Mineur et al. 2015). Most of these species actually fail to establish viable populations, and if persisting, their range in the Mediterranean Sea remains mostly restricted to the lagoon system. Differences in the abiotic physico-chemical environment between the Atlantic and Mediterranean likely underlie the failure of these species to spread widely in the Mediterranean. Nevertheless, repeated observations of Atlantic species in Mediterranean lagoons are evidence for continuous transfers of aquaculture livestock.

The Southeast Mediterranean accounts for 24 first reports, 58% between 1940 and 1960, and a total of 32 introduced species. The construction of the Suez canal in 1896 resulted in an open connection, between the northern Red Sea and the Eastern Mediterranean. As a result, 493 marine species are believed to have invaded the Mediterranean Sea through the Suez canal, so-called Lessepsian or Erythrean migrants (Zenetos et al., 2012). With respect to nonindigenous seaweeds many species were first reported in a series of papers by the Egyptian phycologist Anwar Aleem (1948, 1950). Recent efforts by Greek, Israeli and Turkish phycologists have expanded the list of Lessepsian seaweeds considerably and importantly have confirmed the identity of many species with molecular markers. Nevertheless, a paucity of baseline data makes it often difficult to establish the Lessepsian origin of many species or to point to the exact date of introduction. As outlined above reports of species with clear Indo-Pacific affinities which predate the opening of the Suez canal still puzzle phycologists. In addition the identity of many species reported for the first time by Aleem (e.g. *Gracilaria arcuata, G. disticha, Hypnea flagelliformis*, *Solieria dura, Spatoglossum variabile*) has never been confirmed using molecular markers and is highly uncertain. In general, the lack of solid baseline data hamper a detailed understanding of past and contemporary temporal dynamics of seaweed introductions in the Eastern Mediterranean Sea. More than in any other European region it remains difficult to link the observation of a new seaweed species with the

introduction date. This uncertainty bears down on the monitoring of migration through the Suez canal which is regarded as an ongoing process until present (Boudouresque, 1999). The current construction of the new Suez canal, doubling the capacity of the current corridor, is expected to increase the influx of Red Sea species (Galil et al., 2015) and contribute to the further tropicalization in the Mediterranean Sea (Bianchi, Carlo & Morri, 2003; Coll et al., 2010).

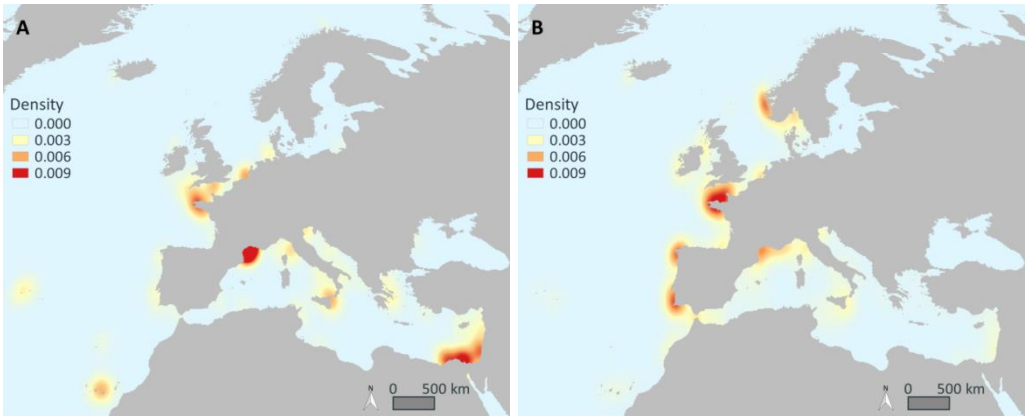


Figure 5. Binned kernel density maps of first introduction events (A) and of the distribution records of introduced species in our database.

Compared to the two Mediterranean hotspots, first reports appear less localized in the NE Atlantic. The Channel (Brittany, southern English coast) and the Scheldt estuary (the Netherlands) are most prominent as introduction hotspots. There is a high correlation between the introduction hotspots and the density map of all records of introduced species, indicative for high monitoring activities in areas where a lot of nonindigenous species are found (Fig. 5B). To some extent this spatial pattern may be influenced by the distribution of phycologists and research institutes. However, there is definitely not a one-on-one relationship between the density map of first reports and the map of all distribution records. Most strikingly, the Eastern Mediterranean Sea (Egypt, Israel) is a clear hotspot for first reports due to their proximity to the Suez Canal, but the total number of records from that region is rather low compared to Atlantic European coasts. On the opposite end of the spectrum, the southern Norwegian coast is particularly well-monitored even though no first records from that area have been reported.

Origin and spread

We mapped the distribution records of introduced seaweed species according to their presumed geographic origin or native range. Distribution records were gridded on a

100 x 100 km raster. Maps depict the number of species per grid cell for species of Northwest Pacific origin (Fig. 6A), Lessepsian migrants (Fig. 6B), Australasian origin (Fig. 6C) and the Northeast and Western Atlantic origin (Fig. 6D). The NW Pacific origin of 45 species is well established (Table 1). An additional 31 species are possibly native to the NW Pacific but there is no strong evidence at present (e.g. molecular sequence data) which support such a claim. Restricting analyses to species for which a NW Pacific origin is not contested, these are predominantly present in the NE Atlantic, with the notable but not surprising exception of the Mediterranean lagoon system (Thau, Venice), and spread relatively little in the Mediterranean. Furthermore, most cells in the Mediterranean Sea, for which species native to NW Pacific have been reported, only contain one species with that origin.

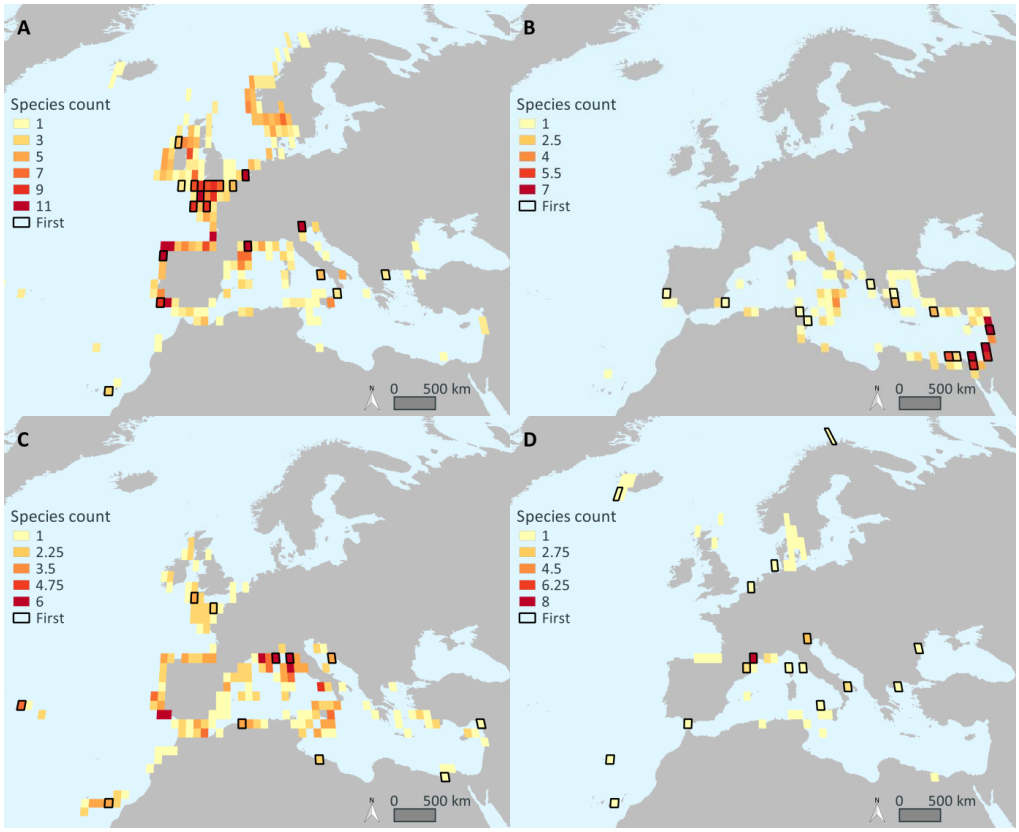


Figure 6. Number of species found in 100 km² grid cells split up by species origin: Northwest Pacific (A), Lessepsian (B), Australasia (C) and Northeast and Western Atlantic (D). Cells containing the first record of one or more species are outlined in black.

In contrast, 34 species with a Lessepsian origin are predominantly distributed in the Eastern Mediterranean Sea with a minority permeating into the Western Mediterranean Sea (Fig. 6B). In contrast 11 species with presumed Australasian origin are predominantly restricted to the Western Mediterranean Sea, Macaronesia and the Atlantic coasts of the Iberian peninsula. Species with Australasian origin appear virtually absent north of Brittany, France. Despite this pattern the introduction vectors for this category of species remains the most elusive. For *Acrothamnion preissii* and *Womersleyella setacea* ship traffic has been suggested as vector based on their first observation close to a major harbour (Livorno, Italy), but accidental release from aquaria is also a possibility (Verlaque et al. 2015). Complicating identification of vectors even further, molecular studies on several nonindigenous species have unveiled multiple independent introductions possibly involving different vectors (McIvor et al., 2001; Provan et al., 2004).

Based on the cumulative plots of the number of species for the most prevalent origins (Fig. 7), we see different patterns depending on the origin of the species. For the NW Pacific we see a sharp increase in the number of first reports around 1970 which slows down after the 1990's. After a big jump in the introduction of species with a Lessepsian origin around 1950, the number of newly reported introduced species has slowly but steadily increased. For the species with an Australasian origin we see that a smaller number of species is being introduced.

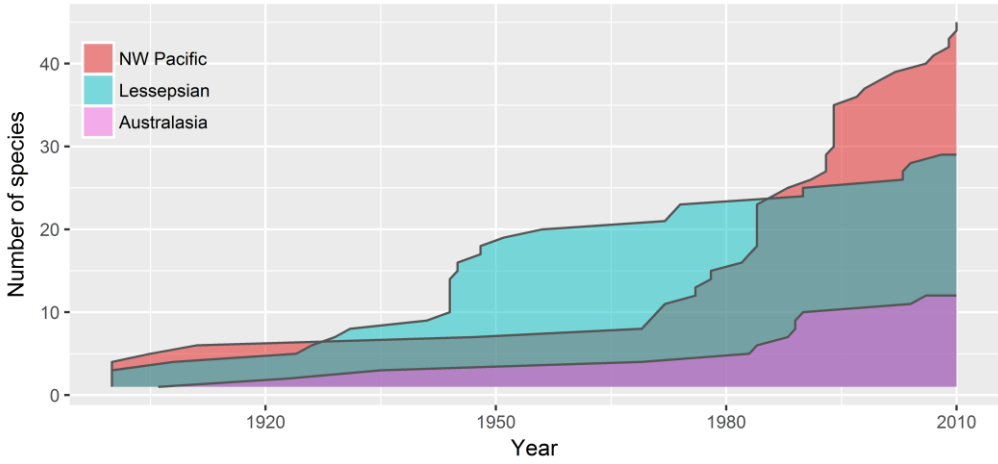


Figure 7. Superimposed cumulative plot of the number of introduced species through time for the most prevalent origins: Northwest Pacific (red), Lessepsian (blue) and Australasia (purple). Only species with a high degree of confidence in their origin are included in this plot.

Conclusion

Detailed analyses of spatial and temporal trends of nonindigenous seaweeds in Europe reveal a complex pattern which can be best understood in terms of the native regions of the species and associated vectors. We identified three different native regions which are responsible for the majority of the nonindigenous species in Europe, the NW Pacific, Australasia and the tropical Indo-Pacific Ocean. Distribution maps of first introductions reveal a non-random pattern with NW Pacific species predominantly introduced in the NE Atlantic region and in lagoon systems in the Mediterranean Sea (Thau and Venice lagoon). Analyses of all distribution records reveal that these species generally do not spread widely in the Mediterranean Sea, but secondary introductions, aided by shellfish transfers from the Atlantic to the Mediterranean lagoon systems and vice versa, are commonly observed. Tropical Indo-Pacific species, predominantly introduced in the Eastern Mediterranean Sea through the Suez Canal remain largely restricted to the latter region with a minority of species spreading to the Western Mediterranean Sea. These species are virtually absent from the Atlantic coasts. The distribution of Lessepsian species likely reflects the environmental tolerance of species with tropical affinities, although one cannot rule out that their current ranges may still expand westward in the Western Mediterranean basin or even Atlantic coasts. Regardless, the distribution of Lessepsian species contrasts to species with Australasian origin who are much more scattered over the entire Mediterranean Sea. Interestingly, Australasian species cannot be easily linked to a specific vector. Fouling, ballast waters and aquarium escapees have all been suggested as vectors (Verlaque et al., 2015). Perhaps the possibility that multiple vectors are involved in the introduction of Australasian species results in the erratic pattern of first reports.

Trends of first reports since 1950 demonstrate that the overall rate of introductions of nonindigenous species is slowing down in Europe. Here we discuss the plausibility of several non-mutually exclusive explanations that could account for the observed decrease in the rate of seaweed introduction in Europe. The most intuitive and optimistic explanation would be that indeed less species have become introduced in Europe during the last two decades. In other words, the measures taken by local and European governments to reduce the influx of nonindigenous species prove effective. The fact that the decline can be attributed primarily to NW Pacific algae (Fig. 7), would corroborate this hypothesis. Livestock transfer of shellfish, the primary vector of algae with a NW Pacific origin, is in principle easier to control compared Lessepsian migration. However, a decline in the rate of reported nonindigenous species doesn't necessarily imply a decrease in introduction rate. Relationships between

introductions and reports of introductions are unfortunately more complicated (Costello & Solow 2003). From the data at hand we can rule out that less attention by the scientific community underlies the decrease in first reports. At least up to the year 2000 the number of records, publication and individual authors showed no sign of decline, while the rate of first reports dropped since 1990. However, it remains possible that a lack of attention in the early second half of the 20th century resulted in a large pool of nonindigenous species waiting to be discovered. If so, the high rates of reports from 1970-1990 could reflect increased scientific interest more than they would reflect introduction rates. The base rate of introductions may have remained constant since 1950, and the pattern of first reports simply reflect a combination of the ease to recognize them and the incentive to report them. A lack of systematic surveys across Europe precludes one from ruling out this scenario. However, there are some indirect indications that introductions rates have not remained constant over the last 50-70 years. Most convincingly, Mineur et al. (2014) correlated Japanese oyster production and disease outbreaks to reports of introduced species in Europe. In addition the difference between Lessepsian and Australasian species which display more constant rates of first reports compared to NW Pacific species is difficult to explain under a constant introduction rate. There is no reason why NW Pacific species would be easier to detect or vice versa.

The observation that at least one source of introductions of marine species in Europe can be controlled, contrasts to the global pattern reported by Seebens et al. (2017) who report an increase across taxonomic groups and geographic regions. Given that livestock transfer of shellfish is the principal vector for the introduction NW Pacific species, it appears that European directives which authorize all movements inside Europe and restrict shellfish stock imports from outside Europe successfully mitigate the influx of nonindigenous species.

While compiling the list of nonindigenous species in Europe, it was quite surprising to encounter so much uncertainty in the primary data at several levels. First, there is taxonomic uncertainty which is rife across the entire geographic region but perhaps even more common in the Mediterranean Sea and Macaronesia. Second, there is also uncertainty as to whether a species is native or introduced in Europe. Both types of uncertainty can be linked but this is not necessarily the case. It is for example possible that a certain species is introduced beyond reasonable doubt, but that the taxonomy is not developed enough to attach a species name. Vice versa, there can be uncertainty on the introduced nature of certain species, despite stable taxonomy. Given this, our final nonindigenous species list should be interpreted with care and we acknowledge that several aspects of the data (e.g. xenoticity) are subjective to some

extent and open for interpretation. Future efforts should be directed toward establishing a DNA-based reference system including European species as well as species from the NW Pacific, Red Sea and other likely donor regions. Reducing the uncertainty in the primary data will be beneficial towards future management of introduced species.

Acknowledgements

The research was carried with financial support from the ERANET INVASIVES project (EU FP7 SEAS-ERA/INVASIVES SD/ER/010) and financial, data & infrastructure support provided by VLIZ as part of the Flemish contribution to the LifeWatch ESFRI.

Supporting information

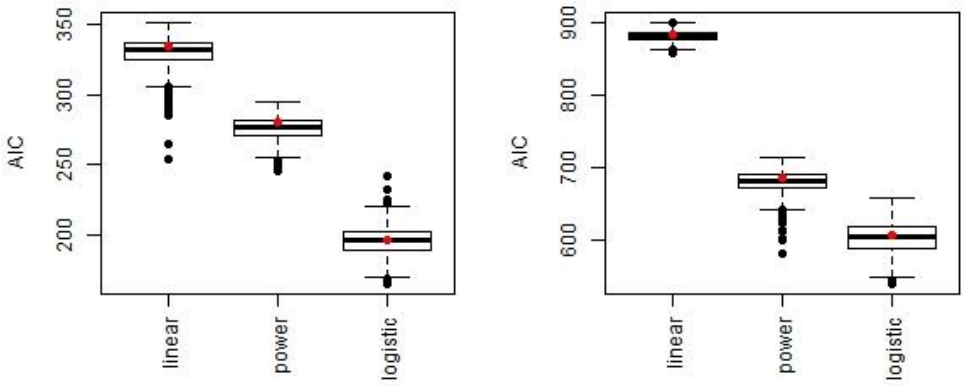


Figure S1. Boxplot of bootstrapped AIC values for curve fitting of the number of introduced species reported for the first time in Europe between 1950 and 2010 (left) and for the number of reported distribution records over all introduced species in the same area and period (right). The fitted curves are a linear curve, a power curve and a logistic curve.

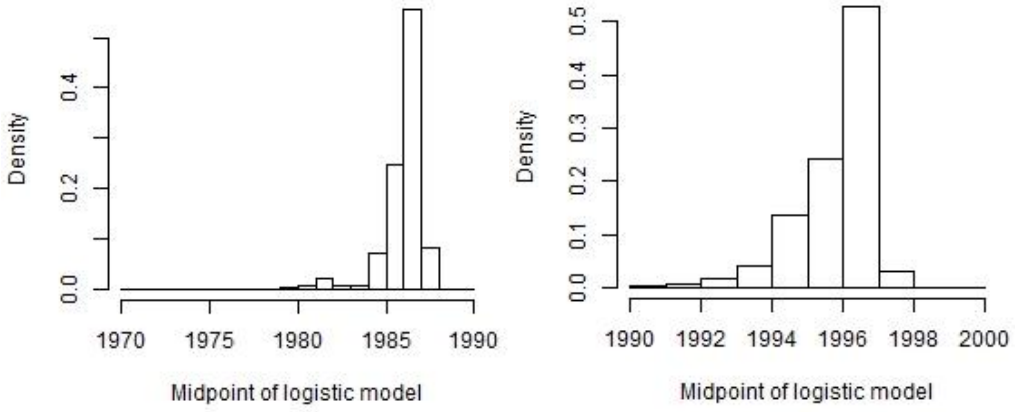


Figure S2. Histograms of the midpoint of the logistic model from 1000 bootstrap samples for the number of introduced species reported for the first time in Europe between 1950 and 2010 (left) and for the number of reported distribution records over all introduced species in the same area and period (right).

Chapter 6

A risk assessment of aquarium trade introductions of seaweed in European waters

Sofie Vranken[1,2], Samuel Bosch[1,2], Viviana Peña[3], Frederik Leliaert[1,4], Frederic Mineur[5] and Olivier De Clerck[1]

[1]*Research Group Phycology, Biology Department, Ghent University, Krijgslaan 281/S8, 9000 Ghent, Belgium*
[2]*Flanders Marine Institute (VLIZ), InnovOcean site, Wandelaarskaai 7, 8400 Ostend, Belgium*
[3]*Universidade da Coruña, BIOCOST Research Group, Facultade de Ciencias, Zapateira, 15071 A Coruña, Spain*
[4]*Botanic Garden Meise, Nieuwelaan 38, 1860 Meise, Belgium*
[5]*School of Biological Sciences, Queen's University, 97 Lisburn Road, Belfast BT9 7BL, United Kingdom*

Submitted in March 2017.

SB contributed the section on the risk of aquarium species.

Abstract

Aquaculture and maritime traffic have been identified as the main vectors for introductions of alien marine species. Except for one notorious case of *Caulerpa taxifolia*, the role of aquarium trade towards the introduction of alien seaweeds has been largely unassessed. Here, we address the risk of accidental release of seaweed species from the aquarium trade market in European waters. We assessed the importance and diversity of seaweed species in the European online aquarium retail circuit. Our web survey revealed more than 30 genera available for online sale into Europe, including known introduced and invasive species. A second aspect of the study consisted in sampling the algal diversity found in various aquaria. While allowing direct and accurate identification of the specimens, this approach was targeting not only ornamental species, but also seaweeds that may be accidentally present in the aquarium circuit. By DNA-barcoding we identified no less than 135 species, of which 7 species are flagged as introduced in Europe with 5 of them reported as invasive. Thermal niche models show that at least 23 aquarium species have the potential to thrive in European waters. As expected by the tropical conditions in most aquaria, southern Atlantic regions of Europe and the Mediterranean are the most vulnerable towards new introductions. Further predictions show that this risk will increase and shift northwards as global warming proceeds. Overall our data indicates that aquarium trade poses a potential but limited risk of new introductions. However, the large reservoir of macroalgal species in aquaria calls for a cautious approach with the highest risk coming from aquaria on in coastal cities and on board of mega yachts.

Introduction

Macroalgae represent one of the largest groups of marine aliens, which may account for 10 to 30% of all marine introduced species in Europe (Schaffelke et al., 2006; Williams & Smith, 2007; Zenetos et al., 2012; Katsanevakis et al., 2013). In areas such as the Thau Lagoon on the French Mediterranean coast, aliens may account for up to one third of the seaweed diversity and up to 100% of the local biomass on hard substrates (Boudouresque et al., 2010). Invasive marine macroalgae may outcompete native biodiversity and affect the functioning of coastal ecosystems (Hammann et al., 2013). For example, *Codium fragile* one of the most hazardous invasive marine macroalgae in temperate regions, is known to outcompete native kelp species (Levin et al., 2002; Scheibling & Gagnon, 2006). Invasions of alien seaweeds do not only pose biodiversity and ecological threats. From an economic perspective, invasive seaweed species may disturb aquaculture and tourism, and eradication and control effort can easily rise to a few million dollars (Neill et al., 2006; Schaffelke & Hewitt, 2007; Irigoyen et al., 2011).

The most important vector for alien seaweeds in Europe appears to be aquaculture and shell fish trade (Zenetos et al., 2012). Indirect evidence, such as the northwestern Pacific origin, time and location of first records, as well as experimental evidence demonstrate the role of oyster transfers as a vector of many seaweed introductions (Mineur et al., 2007a, 2014, 2015). The importance of shellfish transfer as a vector, however, does not imply that other potential pathways are by definition ineffective. Hull fouling or transport by ballast water have been suggested as vectors of invasive species (Hay, 1990; Flagella et al., 2007) but compared to other marine species, these maritime vectors are deemed less important since they exert strong selective pressures. These pressures include the presence of antifouling coatings on ship hulls and the absence of light in non-coated area such as sea chests where heterotrophic fouling organisms can thrive. Moreover, macroalgal propagules do not usually go through a resistant phase that would allow survival or prevent sedimentation in the ballast tanks. As a result, only cosmopolitan opportunistic species are found in standard maritime vectors (Mineur et al., 2007b). Another putative vector is presented by aquarium trade (Padilla & Williams, 2004).

Even though only one introduction, of *Caulerpa taxifolia*, can be ascribed with certainty to aquarium trade (Jousson et al., 1998; Wiedenmann et al., 2001), several other species, including the lionfish *Pterois volitans*, are suspected to have been introduced by accidental releases from aquaria (Whitfield et al., 2002; Zenetos et al., 2012). Some introductions of marine species (*Zebrasoma xanthurum* & *Caulerpa taxifolia*) are even assumed to be caused by accidental release from aquaria on

board mega yachts that travel the world (Meinesz, 1999; Guidetti et al., 2015; Verlaque et al., 2015). Aquarium trade as a pathway for the introduction of marine alien species is, however, still largely unexplored. Moreover, during the last 15 years, the internet has revolutionised how consumers purchase commodities. Trade in living organisms, terrestrial as well as aquatic, forms no exception to this trend. Aquarium hobbyists can obtain assorted living organisms from a wide variety of online sources, ranging from unofficial amateurs to established international suppliers. Recent studies start to point out the importance of biological invasions in aquatic environments associated with online trade (Padilla & Williams, 2004; Walters et al., 2006; Mazza et al., 2015). Most research focuses on freshwater fishes (Rixon et al., 2005; Strecker et al., 2011; Mendoza et al., 2015), the marine seaweed *Caulerpa* (Wiedenmann et al., 2001; Stam et al., 2006; Walters et al., 2006), or on aquarium e-commerce in the USA which is one of the major importers of aquarium species (Padilla & Williams, 2004; Stam et al., 2006; Odom & Walters, 2014). For many other taxa and geographic regions the risk of introducing alien species by aquarium trade remains hitherto unexplored.

The risk of accidental release encompasses not only ornamental species that are directly sold through online or conventional commerce, but also non-target species (i.e. hitchhikers) that can end up in aquarium tanks. One potentially important source for non-target organisms can be found in live rock. Those porous cobbles/boulders are usually pieces of natural reefs (dead scleractininan corals) that have been naturally colonized by a wide range of organisms as coralline and other macro- and microalgae, invertebrates, and bacteria. Such living assemblages not only give the natural look to aquarium reefs that aquarists aspire, but it also serves as a shelter for fishes and invertebrates, as a substrate to sessile organisms, and as biological filtration mechanisms. The popularity of live rock by marine aquarists has been constantly growing since the 1970's (Falls et al., 2008). Unfortunately, live rock also increases the odds of a successful invasion of a wide diversity of species if the aquaria contents are accidentally discharged into the wild. For example, live rock has been reported as a successful vector for jellyfish (Bolton & Graham, 2006).

The present study aims to assess the seaweed diversity currently present in the European aquarium network. To this end, we used two approaches: 1) a surveillance of the online aquarium market for seaweeds that are subject to direct trade, and 2) sampling of aquarium tanks (private, retail shops and wholesalers, and public aquaria) coupled with a DNA barcoding approach, aiming at assessing the total diversity of both traded and accidentally introduced seaweeds. In order to identify the vulnerability of the European regions toward introductions of aquarium-

associated seaweeds, we performed a thermal niche modelling analysis. Since rising temperatures due to climate change are also considered amongst the main threats to biodiversity, these analyses were performed for present and future climate scenarios. To our knowledge, this is the first study that systematically examines the risk of seaweed introductions by aquarium trade extended to total seaweed diversity.

Material and methods

E-trade survey

We monitored the diversity of seaweeds available through e-commerce from August 1 to September 30, 2014. Thereto, we screened online retail and auction sites. Private forums were not monitored because of access restrictions. As similarly done for *Caulerpa* in the US by Walters et al. (2006), a database containing every unique item advertised for sale was compiled, recording the search terms used, vernacular and scientific names mentioned in the advertisement, URL of the commercial site, geographic location of the site, origin of the seaweed, price, availability of information regarding invasive potential, and possibility to ship to Europe. Every online advertisement was saved as a pdf file.

Based on the pictures in the advertisements, we identified all records with best accuracy possible. Every taxon was labelled as 'introduced' or 'not introduced' based on the introduced seaweed distribution maps available on the Seas-era EUPF7ERA-NET INVASIVES projects website (INVASIVES, 2016). 'Introduced' refers to alien species that are directly or indirectly transferred through human activities beyond their natural range of occurrence (Lucy et al., 2016).

Again, We estimated the number of species offered for sale with the incidence-based coverage estimator (ICE), considering every online vendor as a unique sample and the algal species as the diversity. ICE estimates the total species richness by estimating the proportion of the total richness covered by the samples in a set of replicated incidence samples (Gotelli & Colwell, 2010). All calculations were conducted with the program EstimateS 9.1.0 (Colwell & Elsensohn, 2014). Additionally, species accumulation curves were calculated using the R package *vegan* (Oksanen et al., 2017).

Aquarium sampling survey

In order to obtain specimens we contacted associations of aquarists in order to locate owners of ornamental seaweeds and live rocks (i.e. pieces of rock harbouring

a rich variety of microorganisms, invertebrates, and algae collected from tropical reefs), public aquaria, and retail shops. We sampled seaweeds in 5 private aquaria, 4 public aquaria, and 3 retail shops. The identity of the above is not disclosed but can be obtained upon request. We also purchased about 15 live rocks assumed to be originating from Indonesia. We distributed the live rocks in three temperature and light controlled saltwater aquaria and surveyed them for several months. As similarly done with a focus on *Caulerpa* by Walter et al. (2006), we sampled the first seaweeds 4 weeks after the setup, the last after 8 weeks. We preliminarily assigned all the samples to the lowest taxonomic rank possible based on morphology. This resulted in most of cases in an identification to the genus level. We photographed every sample and preserved it in silica gel. Voucher specimens (herbarium and/or formalin preserved) are deposited in the Ghent University Herbarium (GENT). To increase the accuracy of the identifications, we identified the samples by DNA-barcoding. We extracted DNA from silica gel dried specimens with the DNeasy Blood & Tissue kit of Qiagen (Qiagen, Valencia, California, USA) following the manufacturer's instructions. For DNA amplification we followed previously published protocols (McDevit & Saunders, 2009; Saunders & Kucera, 2010; Saunders & Moore, 2013). A complete overview of primers and references is given in Table S1 in Supporting Information. We submitted all the newly generated sequences to Genbank. A complete list of samples and corresponding GenBank accession numbers is provided in Table S2 in Supporting information. PCR products were sequenced by Macrogen. The obtained sequences were aligned with reference sequences from our personal library (Phycology Group, Ghent University) and GenBank with MEGA version 6 (Tamura et al., 2013). We aligned sequences and assigned them to the least inclusive taxonomic rank possible using phylogenetic trees or BLAST searches. Every taxon was again labelled as 'introduced' or 'not introduced' according to the rules described above. Species phylogenetically related to a known introduced species, i.e. belonging to the same genus, were flagged as a 'related'. Asymptotic species richness was estimated with the incidence-based coverage estimator (ICE) using EstimateS 9.1.0 (Chazdon et al., 1998; Colwell & Elsensohn, 2014) and a species accumulation curve was calculated using the R package *vegan* (Oksanen et al., 2017).

Thermal niche

For every unambiguously identified seaweed species, we determined the thermal distribution (i.e. the climatic niche). We used geo-referenced occurrences of the Global Biodiversity Information Facility (GBIF 2016), the OBIS database (OBIS 2016), and published literature sources. To limit the redundancy of neighbouring

occurrence records, we used the Behrmann cylindrical equal-area projection and maintained 1 record per 25 km² grid cell. Secondly, we matched these occurrences to the long-term mean monthly sea surface temperature (SST) values from MARSPEC (Sbrocco & Barber, 2013). After excluding species occurring in less than 30 grid cells, we obtained a data set of 39 species. For each species we calculated the thermal range as the 5th percentile of the SST of the three coldest months and the 95th percentile of the SST of the three warmest months. By using these percentiles as endpoints instead of the minimum and maximum values, we exclude rarities and consider as such the non-static range boundaries of marine species ranges (Bates et al., 2015).

To assess the possible risk of aquarium species to European ecoregions, we tested if the mean SST values of the three coldest and warmest months for a certain European ecoregion were within the thermal range of every aquarium species. If positive, we considered this species as a potential threat for this particular ecoregion. This approximation of habitat suitability was carried out for the current and future (2055) climate. We used the climate model CMIIP5, scenario RCP4.5 (increase of 1.4°C by 2055) of Combal (2014) for vulnerability predictions. The vulnerability of each ecoregion towards new introductions of alien species is estimated as the amount of species that meet the latter rules in that region. The assessed European ecoregions are all ecoregions within the provinces: Northern European Seas, Mediterranean Sea, Black Sea and Lusitanian (Spalding et al., 2007).

Results

E-trade survey

Using 14 different search terms in Google, we identified 39 unique online vendors. The three most successful search terms were ‘Caulerpa for sale uk’, ‘Marine life aquaria’, and ‘Macroalgae aquarium store’. Together, they accounted for more than 50% of the positive hits.

Approximately half of the vendors were professional online retail shops, while the remaining half were online auction pages of hobbyists. Only 1 vendor gave information about the invasive potential of the traded species. The majority of the vendors (27) was situated in the USA. Only one of the US vendors exported to Europe, 16 did not ship to Europe, and 10 did not specify the countries shipped to. Other vendors were located in France, Germany, Malaysia, Poland, Thailand, and the

United Kingdom. These vendors all shipped to or within Europe. Only one vendor gave information on the origin or the invasive potential of species.

In total we estimated the seaweed diversity distributed by the 39 online vendors at 75 species belonging to minimum 53 genera, based on a total of 236 unique sale items (Table 1). The number of species should be considered an underestimation of the true diversity since identification to species level was often not possible based on the limited information provided in the advertisements. Genus-level diversity is therefore more accurate and will be used primarily in the subsequent analyses. The ICE diversity coverage estimator resulted in a total estimated diversity of 123 species and 100 genera based on 39 vendors (Fig. 1). This large number is confirmed by the non-asymptotic nature of the species accumulation curve created for 39 online vendors (Fig. S1 in Supporting information). For three quarter of all online records, species (30%) or genus names (46%) were provided by the vendors, while the remainder did not bear a scientific name. Obvious misidentifications by the vendors at species and genus level occurred, respectively, in 3 and 5% of the cases. Vernacular names ranged from commonly used names like 'sea lettuce' (*Ulva* sp.) to less obvious names like 'dragon's breath' (*Halymenia* sp.) and 'tang heaven' (*Gracilaria* sp.). 60% of the seaweeds available through global e-commerce belonged to the green algae (Chlorophyta), 36% to the red algae (Rhodophyta), and 4% to the brown algae (Phaeophyceae). *Caulerpa*, *Chaetomorpha*, and *Halimeda,* accounted for half the records of Chlorophyta. Within the Rhodophyta, most of the records belonged to *Gracilaria* and *Botryocladia.* Phaeophyceae were hardly offered for sale, and only occasionally *Lobophora*, *Padina* or *Sargassum* was encountered. For 71% of the advertisements it was not possible to ship to Europe, or shipping details were not provided. Only one third of the seaweeds could be purchased in Europe. Biodiversity trends were similar for the European as for the global aquarium trade network with the majority of seaweeds belonging to the Chlorophyta. We found 30 available genera on the European online trade market (Table 1). More than half of the records found on the European e-market belong to genera that include species introduced in Europe. Moreover, several species flagged as invasive, or species closely related to invasive species are offered for sale. On a genus-level 26% of the specimens offered for sale can be classified as invasive or potentially invasive. Invasive species found were *Caulerpa taxifolia* and *C. cylindracea* (often under the name *C. racemosa*). Other species of *Caulerpa*, *Codium*, and *Sargassum* were considered as potentially invasive (Boudouresque & Verlaque, 2002; Streftaris & Zenetos, 2006; Provan et al., 2008).

Table 1. Genera found on the online trade market with their status of introduction in Europe and the number of record available in and outside the European online market. 'introduced' (INT) represents genera that include species introduced in Europe, 'not introduced' (NI) genera that do not include species introduced in Europe, when unclear or unknown the status is represented by 'uncertain' (UNC).

Genus	Status	Number of records (European market)	Number of records (non-European market)	Total
Chlorophyta				
Acetabularia	NI		1	1
Boergesenia	NI	1		1
Bornetella	NI	1		1
Caulerpa	INT	20	32	52
Chaetomorpha	UNC	4	17	21
Chlorodesmis	NI	2	3	5
Cladophora	INT	6	3	9
Codium	INT	1	6	7
Cymopolia	NI		4	4
Enteromorpha	NI		1	1
Halimeda	NI	6	9	15
Neomeris	INT	1	2	3
Penicillus	NI		3	3
Rhipocephalus	NI		2	2
Udotea	NI	1	3	4
Ulva	INT	1	9	10
unknown	UNC		1	1
Valonia	NI	2		2
Rhodophyta				
Acanthophora	INT		3	3
Actinotrichia	NI	1		1
Agardhiella	INT		1	1
Amansia	NI	1		1
Amphiroa	NI	2		2
Amphiroa	INT		3	3
Botryocladia	INT	3	7	10
Bryothamnion	NI		1	1
Carpopeltis	NI		4	4
Ceramium	INT	1		1
Cryptomenia	INT		1	1
Dichotomaria	NI	3		3
Eucheuma	NI		2	2
Fauchea	NI		1	1
Galaxaura	INT	1	4	5
Gracilaria	INT		17	17
Haliptilon	NI	1		1
Halymenia	NI	1	4	5
Heterosiphonia	NI		2	2
Hypnea	INT		1	1
Jania	NI	1		1
Kappaphycus	NI	1		1
Liagora	NI		1	1
Lithothamnion	NI	1		1
Mastophora	NI	1		1

Genus	Status	Number of records (European market)	Number of records (non-European market)	Total
Osmundaria	NI		1	1
Peyssonnelia	NI	1		1
Portieria	NI		4	4
Ptilophora	NI		2	2
Scinaia	NI		1	1
Phaeophyceae				
Canistrocarpus	NI		1	1
Dictyota	INT	1		1
Lobophora	NI		2	2
Padina	NI	1	1	2
Sargassum	INT	1	1	2
Turbinaria	NI	1		1
unknown	UNC		6	6
Total		69	167	236

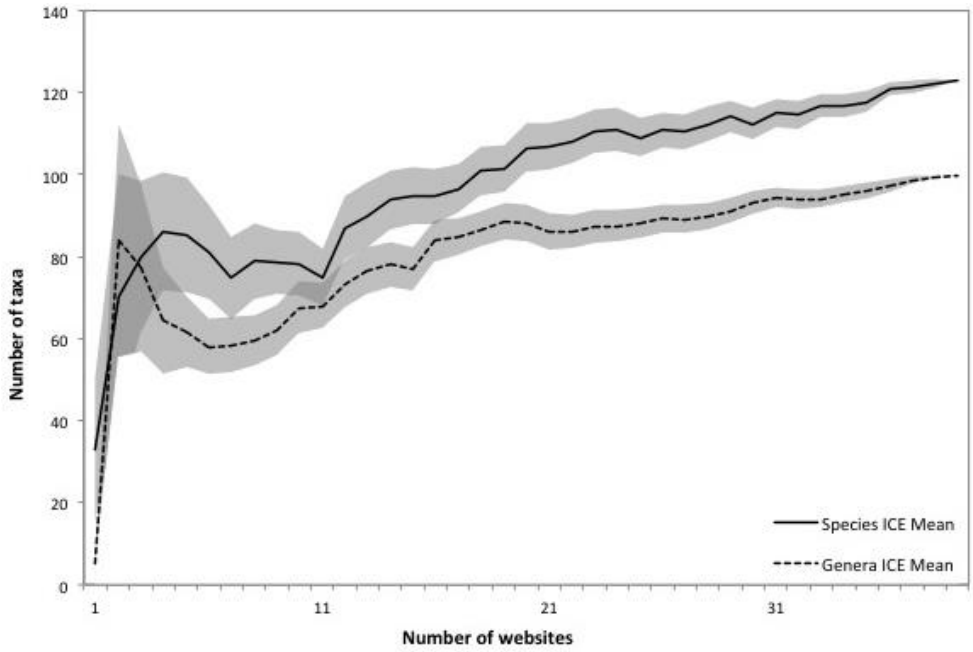


Figure 1. Incidence-based Coverage Estimator (ICE) for species and genera found on the global e-market (mean ± SE).

Aquarium sampling survey

We identified 217 specimens from almost 50 aquarium tanks from private aquaria, public aquaria, and retail shops. Identifications were based on a combination of morphology and DNA barcoding (Table 2). 29 samples were identified to genus level and 189 specimens to species level, of which more than half were assigned to named species. Half of the species not assigned to a named species belonged to the coralline algae (Corallinales). In total, we found 135 unique seaweed taxa (Table 2),

of which almost half belonged to either the Chlorophyta or the Rhodophyta. Only a minority of the samples (4%) belonged to the Phaeophyceae. The Chlorophyta and Rhodophyta were equally sampled in aquarium tanks but the diversity of the Rhodophyta was significantly higher. Especially coralline red algae (subclass Corallinophycidae) were highly divers and abundant; they accounted for 57% of total seaweed diversity found and for 26% of the samples collected. Within the Rhodophyta, the following most abundant genera were *Botryocladia*, *Haraldiophyllum* and *Polysiphonia*. *Caulerpa, Chaetomorpha* and *Cladophora* were the most abundant green algae, and *Dictyota* the most abundant brown alga. The ICE diversity coverage estimator estimates the total diversity on 370 species and 128 genera (Fig. 2). Similar to e-commerce websites, this large number is confirmed by the clearly non-asymptotic nature of the species accumulation curve for the aquarium samples (Fig. S2 in Supporting information). We found 6 species that are known to be introduced in Europe of which 5 species are reported as invasive: *Caulerpa taxifolia*, *Asparagopsis taxiformis*, *Hypnea valentiae*, *Womersleyella setaceae* and *Sargassum muticum* (Table 2) (Boudouresque & Verlaque, 2002; Chualáin et al., 2004; Streftaris & Zenetos, 2006; Provan et al., 2008; Nikolić et al., 2010). Another 40 species were closely related to introduced species. These account for 30% of all specimens sampled in the European aquaria.

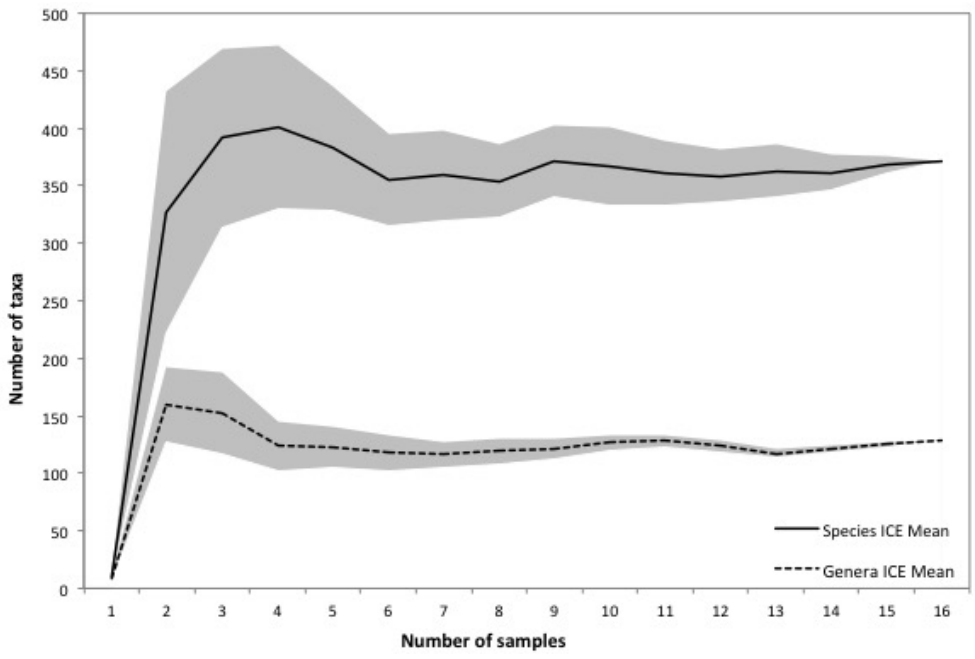


Figure 2. Incidence-based Coverage Estimator (ICE) for species and genera found in the European aquarium trade market (mean ± SE).

Table 2. Seaweed diversity found in the European aquarium network and their status of introduction in Europe. 'not introduced' (NI) indicates species not known to be introduces in Europe, 'introduced' (INT) indicates species reported as introduced in Europe, 'uncertain' (UNC) indicates that the status of introduction is unclear or unknown, 'related' (REL) indicates that a congeneric species is reported as introduced in Europe.

Chlorophyta			**Rhodophyta**			**Phaeophyceae**		
Species	Status	Nr of Records	Species	Status	Nr of Records	Species	Status	Nr of Records
Caulerpa parvifolia	NI, REL	9	*Mesophyllum sp1*	NI	5	*Dictyota friabilis1*	NI, REL	4
Chaetomorpha vieillardii	NI	7	*Haraldiophyllum sp1*	NI	4	*Dictyota ceylanica4*	NI, REL	1
Caulerpa racemosa	UNC	6	*Sporolithon sp1*	NI	3	*Dictyota implexa*	NI, REL	1
Caulerpa constricta	NI, REL	5	*Titanophora sp1*	NI	3	*Halopteris filicina*	NI, REL	1
Caulerpa taxifolia	INT	5	*Acanthophora spicifera*	NI, REL	2	*Sargassum muticum*	INT	1
Cladophora	REL	4	*Acrosymphyton sp1*	NI	2	*Sargassum sp1*	REL	1
Chaetomorpha	UNC	3	*Antithamnion*	REL	2			
Cladophora albida/sericea	NI, REL	3	*Asparagopsis taxiformis*	INT	2			
Derbesia	REL	3	*Botryocladia sp1*	NI, REL	2			
Halimeda gigas	NI	3	*Cryptonemia sp1*	NI, REL	2			
Valonia macrophysa	NI	3	*Gracilaria vieillardii*	NI, REL	2			
Bryopsis	NI	2	*Harveylithon sp1*	NI	2			
Bryopsis sp1	NI	2	*Lithophyllum sp2*	REL	2			
Bryopsis sp3	NI	2	*Melobesioideae sp2*	NI	2			
Caulerpa cupressoides	NI, REL	2	*Peyssonnelia japonica*	NI	2			
Caulerpa prolifera	NI, REL	2	*Peyssonnelia sp3*	NI	2			
Caulerpa sertularioides	NI, REL	2	*Polysiphonia*	REL	2			
Cladophora herpestica	INT	2	*Polysiphonia sp1*	NI, REL	2			
Cladophora pellucida	NI, REL	2	*Ramicrusta sp1*	NI	2			
Cladophora prolifera	NI, REL	2	*Sporolithon sp3*	NI	2			
Derbesia sp3	NI, REL	2	*Yonagunia zollingeri*	NI	2			
Halimeda minima	NI	2	*Amphiroa*	NI	1			
Valonia utricularis	NI	2	*Asparagopsis*	REL	1			
Boergesenia forbesii	NI	1	*Botryocladia*	REL	1			
Boodlea sp1	NI	1	*Botryocladia sp2*	NI, REL	1			
Boodlea sp13	NI	1	*Ceramium codii*	NI, REL	1			

Chlorophyta			**Rhodophyta**			**Phaeophyceae**		
Species	Status	Nr of Records	Species	Status	Nr of Records	Species	Status	Nr of Records
Boodlea sp2	NI	1	*Ceratodictyon repens*	NI	1			
Bryopsis sp2	NI	1	*Chondracanthus saundersii*	NI, REL	1			
Caulerpa chemnitzia	NI, REL	1	*Coelarthrum*	NI	1			
Caulerpa flexilis	NI, REL	1	*Crouania attenuata*	NI	1			
Caulerpa lentillifera	NI, REL	1	*Cryptonemia lomation*	NI, REL	1			
Caulerpa oligophylla	NI, REL	1	*Erythrotrichia carnosa*	NI	1			
Caulerpa serrulata	NI, REL	1	*Griffithsia sp1*	NI, REL	1			
Chaetomorpha sp1	UNC	1	*Halymenia durvillei1*	NI	1			
Chaetomorpha sp2	UNC	1	*Halymenia durvillei2*	NI	1			
Chaetomorpha sp3	UNC	1	*Hydrolithon sp1*	NI	1			
Chlorodesmis	NI	1	*Hydrolithon sp2*	NI	1			
Cladophoropsis	REL	1	*Hydrolithon sp3*	NI	1			
Codium	REL	1	*Hypnea sp1*	NI, REL	1			
Codium arenicola	NI, REL	1	*Hypnea valentiae*	INT	1			
Codium dwarkense	NI, REL	1	*Incendia sp1*	NI	1			
Derbesia sp1	NI, REL	1	*Laurencia sp1*	NI, REL	1			
Derbesia sp4	NI, REL	1	*Lithophyllum sp1*	REL	1			
Halimeda disoidea	NI	1	*Lithophyllum sp3*	REL	1			
Halimeda opuntia	NI	1	*Lithophyllum sp4*	REL	1			
Parvocaulis parvula	NI	1	*Lithophyllum sp5*	REL	1			
Ulva	REL	1	*Mastophoroideae sp1*	NI	1			
Ulva laetevirens	NI, REL	1	*Mastophoroideae sp2*	NI	1			
Ulva sp1	NI, REL	1	*Melobesioideae sp1*	NI	1			
Ulva sp2	NI, REL	1	*Meredithia sp1*	NI	1			
Ulvella	NI	1	*Mesophyllum sp2*	NI	1			
Ulvella leptochaete	NI	1	*Mesophyllum sp3*	NI	1			
			Mesophyllum sp4	NI	1			
			Neosiphonia sp1	NI, REL	1			
			Palisada sp1	NI	1			
			Peyssonnelia sp1	NI	1			

Chlorophyta			**Rhodophyta**			**Phaeophyceae**		
Species	Status	Nr of Records	Species	Status	Nr of Records	Species	Status	Nr of Records
			Peyssonnelia sp2	NI	1			
			Peyssonnelia sp4	NI	1			
			Peyssonnelia sp5	NI	1			
			Peyssonnelia sp6	NI	1			
			Peyssonnelia sp7	NI	1			
			Phymatolithon sp1	NI	1			
			Plocamium sp1	NI, REL	1			
			Pneophyllum	NI	1			
			Polystrata sp1	NI	1			
			Porolithon sp	NI	1			
			Pterocladiella caerulescens	NI	1			
			Pterocladiella sp1	NI	1			
			Ptilophora scalaramosa	NI	1			
			Rhodymenia ardissonei	NI, REL	1			
			Rhodymeniaceae	NI	1			
			Sarconema filiforme	INT	1			
			Sarconema sp1	NI, REL	1			
			Sporolithon sp2	NI	1			
			Titanoderma sp1	NI	1			
			Womersleyella setacea	INT	1			
			Yonagunia sp1	NI	1			
Total		104	Total		105	Total		9

Thermal niche

Comparison of the thermal distribution of the aquarium species with the current temperature conditions demonstrated that at least 23 of these species could possibly thrive in European seas under current climate conditions. This number increases to minimum 26 species in 2055 under future climate change scenario CMIIP5, RCP4.5. The majority of these species is already present in Europe and not known to be invasive (Table 3). Following our predictions, the number of aquarium seaweed species that is able to survive in the European waters is higher for the warmer southern European regions than for the northern, cooler ecoregions. The Aegean Sea, the Levantine Sea and the Saharan Upwelling were suitable for at least 12 more species than presently reported (Fig. 3A). When only species known to be introduced are considered, 4 more introduced species could thrive in the ecoregions Azores Canaries Madeira, Ionian Sea and Saharan Upwelling under the current climate (Table 3). Extrapolating predictions to the climate predicted in 2055 under CMIIP5, RCP4.5 reflects a northward trend in invasion risk (Fig. 3B). All species considered are estimated to be able to thrive in more ecoregions under future climate conditions (2055) then under actual and estimated current (2010) conditions (Table S3 in Supporting Information). The Adriatic Sea (+7 species), the Baltic Sea (+4 species), the Black Sea (+4 species) and the South-European Atlantic Shelf (+4 species) had the biggest increase in invasion risk (Fig. 3B).

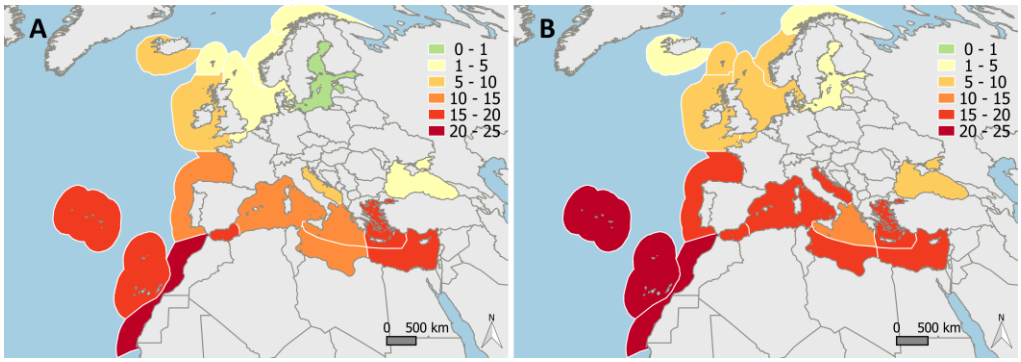


Figure 3. The risk of new introductions by aquarium seaweed species in Europe estimated by the number of species with a thermal distribution falling within the mean maximum and minimum SST for each ecoregion under current (A, 2010) and future (B, 2055) climate conditions (model CMIIP5 scenario RCP4.5).

Table 3. Number of aquarium species found (actual records) and estimated under current and future (2055) climatic conditions for all European ecoregions. Between brackets are the number of species that are known to be introduced in Europe or (/) in another part of the world.

Ecoregion	Actual records	Current climate	Future (2055)
Adriatic Sea	4 (1/1)	9 (0/1)	16 (4/1)
Aegean Sea	4 (2/0)	16 (4/1)	16 (5/1)
Alboran Sea	11 (2/1)	17 (5/1)	19 (5/2)
Azores Canaries Madeira	13 (2/1)	20 (6/2)	21 (6/2)
Baltic Sea	2 (0/1)	0 (0/0)	4 (1/1)
Black Sea	2 (0/0)	4 (0/0)	8 (0/1)
Celtic Seas	9 (1/2)	10 (1/2)	10 (1/2)
Faroe Plateau	2 (0/1)	5 (1/1)	7 (1/1)
Ionian Sea	3 (1/0)	14 (5/1)	15 (5/2)
Levantine Sea	4 (4/0)	16 (5/2)	17 (6/2)
North Sea	7 (1/1)	5 (1/1)	7 (1/1)
Northern Norway and Finnmark	0 (0/0)	2 (0/1)	4 (1/1)
Saharan Upwelling	6 (2/1)	22 (6/3)	23 (6/3)
South and West Iceland	2 (0/1)	6 (1/1)	5 (1/1)
South European Atlantic Shelf	10 (2/1)	14 (3/2)	18 (5/2)
Southern Norway	4 (1/1)	5 (1/1)	6 (1/1)
Tunisian Plateau/Gulf of Sidra	5 (3/1)	14 (4/2)	17 (6/2)
Western Mediterranean	16 (4/1)	15 (4/1)	17 (5/1)
Europe	21 (7/2)	23 (7/3)	26 (7/4)

Discussion

The risk posed by aquarium trade as a vector for introductions of alien aquatic taxa has relatively recently been raised and demonstrated by several studies (Padilla & Williams, 2004; Rixon et al., 2005; Walters et al., 2006; Mazza et al., 2015; Howeth et al., 2016). The vast majority of these studies focus on freshwater species and the USA which is considered as one of the major importers of aquarium species of the world (Padilla & Williams, 2004). Our survey confirms that online aquarium trade in marine macroalgae is best established in the USA. Only a minority of the online vendors ship to or in Europe, which limits the possible risk of introductions of aquarium associated introductions in Europe substantially. Despite the smaller market share, the seaweed diversity offered on the European e-market is, nevertheless, almost as high as the diversity on the non-European market. We found 75 species available online of which 30 could be shipped in or to Europe. Only one third of the species is advertised on both the European and the non-European e-market.

Aquarists often purchase or exchange organisms informally, in aquarist clubs, or through internet forums (personal communication aquarists). Since these purchasing alternatives are very hard to monitor and not considered in this study, the marine aquarium related diversity remains partly unexplored. Furthermore, these informal pathways will be very hard to regulate with respect to management strategies. Important is that 26% of the macroalgae offered for sale online are flagged as potentially invasive which creates a realistic risk for possible new hazardous introductions. Previous research has proven that *Caulerpa* is an important player of the aquarium trade in the United States (Stam et al., 2006; Walters et al., 2006). But invasive *Caulerpa* strains are rarely encountered on the American e-market, most likely due to awareness campaigns and legal regulation on trade of *C. taxifolia (Stam et al., 2006; Walters et al., 2006)*. These authors recommend, however, a full ban of the *Caulerpa* genus due to the poor identification of traded algae (which is confirmed by our results), the need of molecular tools to identify invasive strains, and the lack of understanding of the potential invasive capacity of other *Caulerpa* species (Stam et al., 2006; Walters et al., 2006). Our survey indicates that also in Europe *Caulerpa* is by far the most common genus offered for sale online (Table 1). Corresponding to Mazza et al. (2015) we also found *Caulerpa taxifolia* online, confirming the potential dispersal of this invasive species through aquarium e-commerce and illustrating the need of legal restrictions regarding online aquarium trade of macroalgae in Europe. A few cases were identified where tropical seaweeds collected in their natural environment (Malaysia and Thailand) are offered for sale online, thereby increasing the risk of introducing new potentially invasive species. We found no information about the treatment of the shipped seaweed material. Therefore, also inconspicuous organisms attached to the shipped seaweed material or present in the shipping water may be transported. Furthermore, this trade of newly collected specimens would also increase the genetic diversity within aquarium traded and potentially introduced seaweed species and other organisms.

We identified minimum 135 taxa in the private and public aquaria, and retail shops. The number of estimated taxa reached a plateau (Fig. 2), which is indicative for a representative sampling. Identification of seaweed species based on morphological features is not straightforward, and therefore DNA sequence data are used to guide species identification (DNA barcoding) (Saunders, 2005; Leliaert et al., 2014). Although DNA barcoding has proven effective for rapid species identification in algae, an important limitation is the lack of a comprehensive DNA-based reference framework. This is especially the case for the coralline red algae, a group comprising a large part of unresolved biodiversity. Despite this difficulty identifying species, we

identified 85% of the 217 samples to species level based on molecular data. This shows that aquaria host substantial unknown diversity.

Like the available online seaweed diversity, the diversity sampled in aquaria was highest for Rhodophyta. This high diversity in Rhodophyta is mainly due to the high abundance of coralline red algae (44 species). These calcified algae are popular among aquarists because of their appealing colour and good covering of the tank. Therefore, aquarists often add supplements to enhance growth of coralline algae (personal communication aquarists). Chlorophyta are popular among aquarists as biological filtration mechanism (e.g. *Caulerpa*, *Chaetomorpha*) (Odom & Walters, 2014). Popular macroalgae, such as *Bortryocladia, Chaetomorpha, Caulerpa*, are easily maintained in aquarium conditions because they have broad environmental tolerances, exhibit rapid growth, vegetative reproduction and high reproductions rates. These are also characteristics linked to invasive seaweeds (Thomsen & McGlathery, 2007; Andreakis & Schaffelke, 2012). A worrying concern emerging from our survey is the presence of introduced and known invasives or species related to invasives, including *Caulerpa taxifolia*, *Asparagopsis taxiformis* and *Womersleyella setacea*. Aquarium associated species may therefore pose a realistic threat to European coasts.

The diversity found in the sampled aquaria is remarkably larger than the diversity found online. Species found online are mostly large species used for ornamental purposes, fish food, or to a lesser extent, filtration purposes, while the diversity samples in the aquaria also includes small, epibiotic species that are often accidentally introduced in the aquaria through other organisms or live rocks. Especially live rocks prove to be a successful vector for a variety of species (Bolton and Graham 2006; Walters et al. 2006; this study). Walters et al., (2006) mentioned the development of 25 seaweed species, next to 4 *Caulerpa* species from live rock. Several genera we observed (e. g. *Caulerpa*, *Hydrolithon, Peyssonnelia, Dictyota, Cladophoropsis,* and *Valonia*) were already recorded to develop from live rock by Fosså & Nilsen (1996). Furthermore, we observed polychaetes, hydroids and cyanobacteria developing from the live rocks. These specimens have not been further surveyed but this highlights that live rock is a successful vector for an unknown variety of organisms, including inconspicuous microorganisms. Next to tropical seaweed species we found in warm water aquaria, we also found European species in cold water aquaria (e.g. *Dictyota implexa, Halopteris filicina, Cladophora albida*). These examples were the result of private samplings by the responsible of the aquarium (personal communication). This indicates that aquarists also acquire

seaweeds through informal ways and in this case even facilitates intra-European introductions.

The estimated asymptotic species richness was both for the e-trade as well as the aquaria far larger than the number of species identified indicating that there is relatively large remaining diversity to be uncovered (Figs. 1 & 2). This was confirmed by the species accumulation curves

Comparison of the mean SST and temperature range of the aquarium species demonstrates that European aquarium trade may not pose an imminent risk towards introductions of new macroalgae in European ecoregions. Most of the species are either already established in Europe or are not able to thrive in European ecoregions. But additional introductions may however result in an expansion of the genetic diversity of these invasive species. The higher risk of introduction in the southern parts of Europe is to be expected, as most species found in the aquaria are tropical species. As climate change proceeds, most ecoregions will become suitable to a higher number of aquarium species (Fig. 3 & Table 3). The invasive species included in the risk assessment (*Asparagopsis taxiformis, Caulerpa taxifolia, Sargassum muticum, Womersleyella setacea*) are all able to thrive in more ecoregions after climate change then under current conditions (Table S3). Note that while a thermal range of a species may not fully overlap the thermal range of an ecoregion, there might be smaller parts of that ecoregion that are suitable for a species. Consequently, the estimated number of species that can thrive in an ecoregion may be higher than we calculated. Conversely, given that only temperature was used to estimate the introduction risk, other factors restricting the distribution of macroalgae such as salinity and substrate may render specific ecoregions less suitable. We expect this to be especially the case for the Baltics and the Black Sea as they have a very specific salinity profile. These findings support the hypotheses of Rixon et al. (2005) that the probability of aquarium species establishment along European coasts will increase with climate warming because most aquarium species are of tropical or subtropical origin.

Eradication of invasive species once they are established is very challenging. Hence prevention of new introductions is most effective in avoiding and limiting new biological invasions (Doelle et al., 2007; Vander Zanden & Olden, 2008). Research like this study, that focuses on identification of possible vectors of invasive species geographic regions and ecosystems most susceptible to them, is therefore essential in the development of effective management strategies (Stam et al., 2006; Vander Zanden & Olden, 2008; Corriero et al., 2016). Although global awareness regarding

invasive species is growing, the development of legal restrictions is slow. The European Union has recently developed a blacklist of species for which keeping, importing, selling, breeding, and growing are restricted. This list contains only 37 species (mostly marine and terrestrial animals, and land plants), and no macroalgae (European Parliament, 2014; European Commission, 2016). The trade of macroalgal species is not restricted by CITES regulations, but the trade of live rocks is (CITES, 2006).

It has been previously stated that the probability of introduction of aquarium species is higher in regions close to large coastal cities and in regions where mega yachts with on-board marine aquaria are common due to a higher chance of transfer of seaweed material to the sea (Johnston & Purkis, 2014; Guidetti et al., 2015). Personal communication with aquarists revealed that many aquarists dispose their waste in ways that should prevent future introductions; i.e. putting waste in solid waste for landfill or solid waste for compost, which is encouraging. There were unfortunately also aquarists that dump their aquarium waste in the indoor plumbing or garden (personal communication), which may be dangerous in regions in close vicinity of the coast. Adding bleach to or boiling waste before dumping are possible solutions to avoid new introductions. Next to trade related legislations, proper education of aquarists has proven to help to prevent new introductions (Padilla & Williams, 2004; Walters et al., 2006) and is welcome here. But to fully eliminate the introduction risk by aquarium trade, policy-making bodies should further legal restrictions.

Acknowledgements

This work was financially supported by Ghent University through the project "Towards integrated European marine research strategy and programmes - SEAS-ERA" (ERAC-CT2009-249552) within the framework of the EU ERA-Net initiative (7th Framework Program). V. P. acknowledges support from the postdoctoral programs Campus Industrial de Ferrol (Universidade da Coruña) and Plan I2C (Xunta de Galicia).

Supporting information

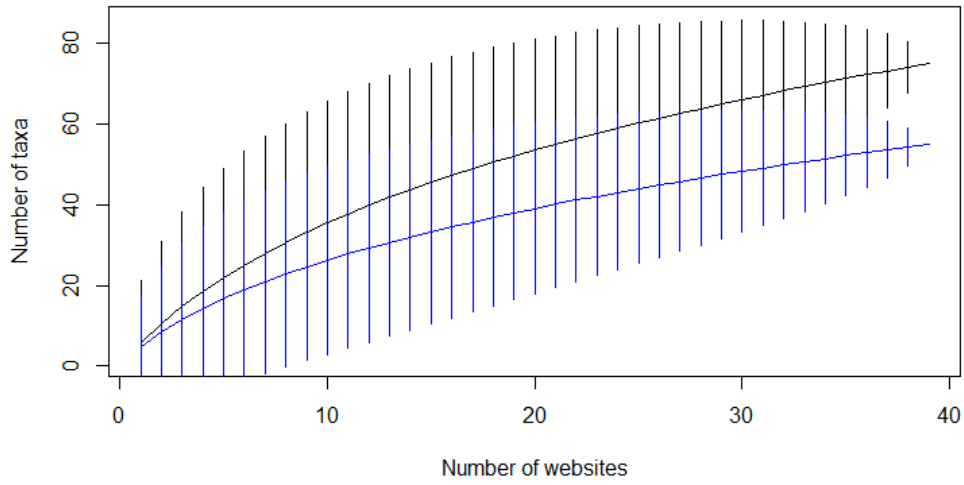


Figure S1. Species accumulation curves for the number of species (black) and genera (blue) found in the e-commerce websites, each website represents one sample event and the vertical bars represent the standard deviation.

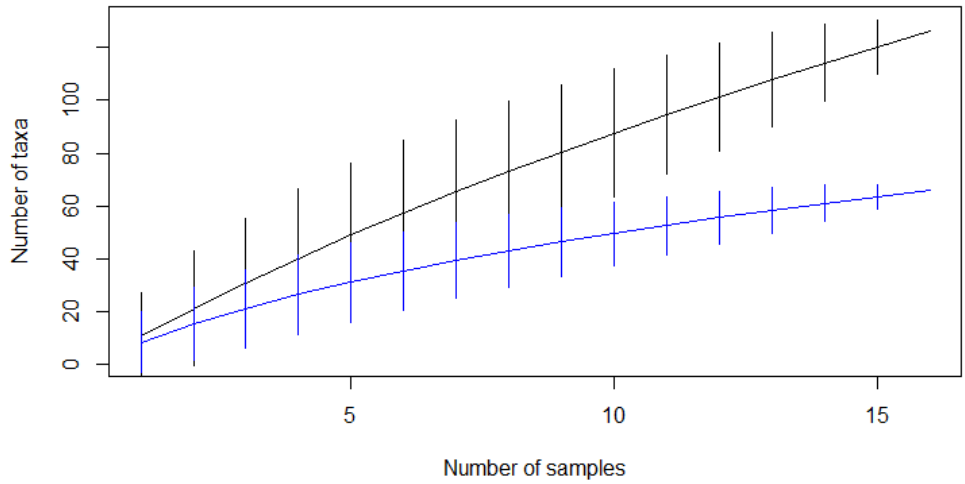


Figure S2. Species accumulation curves for the number of species (black) and genera (blue) found in the 16 public and private aquaria and retail shops, each aquarium representing one sample event and the vertical bars representing the standard deviation.

Table S1. Primers used for PCR amplification and sequencing.

	Forward primer	Reverse primer	Reference
Chlorophyta			
ITS1	TW3	H1R	(Leliaert et al. 2009)
ITS2	TW5	ITS4	(Leliaert et al. 2009)
RBCL1	7F	712F	(Verbruggen et al. 2009)
TUFA1	Tuf AF	Tuf AR	(Verbruggen et al. 2009)
LSU	C'1FL		(Leliaert et al. 2007)
SSU	SR1 SSU897	SS11H 18Sc2	(Bakker et al. 1994; Hanyuda et al. 2002; Leliaert et al. 2007)
Phaeophyceae			
COX1	COX1F_Dic		(Tronholm et al. 2010)
PsbA	psbAF1		(Yoon et al. 2002)
Rhodophyta			
RBCL2	F8 F481		(Draisma et al. 2001)
PSBA	psbAF1		(Yoon et al. 2002)

Bakker FT, Olsen JL, Stam WT, Van Den Hoek C. 1994. The cladophora complex (Chlorophyta): new views based on 185 rRNA gene sequences. Molecular Phylogenetics and Evolution **3**:365-382.

Draisma SG, Prud'Homme van Reine WF, Stam WT, Olsen JL. 2001. A reassessment of phylogenetic relationships within the Phaeophyceae based on RUBISCO large subunit and ribosomal DNA sequences. Journal of Phycology **37**:586-603.

Hanyuda T, Wakana I, Arai S, Miyaji K, Watano Y, Ueda K. 2002. Phylogenetic relationships within Cladophorales (Ulvophyceae, Chlorophyta) infered from 18S rRNA gene sequences, with special reference to Aegagropila linnaei1. Journal of phycology **38**:564-571.

Leliaert F, De Clerck O, Verbruggen H, Boedeker C, Coppejans E. 2007. Molecular phylogeny of the Siphonocladales (Chlorophyta: Cladophorophyceae). Molecular phylogenetics and evolution **44**:1237-1256.

Leliaert F, Verbruggen H, Wysor B, De Clerck O. 2009. DNA taxonomy in morphologically plastic taxa: algorithmic species delimitation in the Boodlea complex (Chlorophyta: Cladophorales). Molecular Phylogenetics and Evolution **53**:122-133.

Tronholm A, Steen F, Tyberghein L, Leliaert F, Verbruggen H, Antonia Ribera Siguan M, De Clerck O. 2010. Species Delimitation, Taxonomy, and Biogeography of Dictyota in Europe (Dictyotales, Phaeophyceae) 1. Journal of phycology **46**:1301-1321.

Verbruggen H, et al. 2009. A multi-locus time-calibrated phylogeny of the siphonous green algae. Molecular Phylogenetics and Evolution **50**:642-653.

Yoon HS, Hackett JD, Pinto G, Bhattacharya D. 2002. The single, ancient origin of chromist plastids. Journal of Phycology **38**:40-40.

Table S2. List of specimens sampled in aquaria.

Taxon	Sample ID	Location
Caulerpa racemosa	SV0001	Live rock 1
Boodlea sp13	SV0002	Live rock 1
Boergesenia forbesii	SV0003	Live rock 1
*Chaetomorpha vieillar*dii	SV0004	Live rock 1
Chaetomorpha	SV0005	Live rock 1
Chaetopmorpha sp1	SV0006	Live rock 1
Chaetopmorpha sp2	SV0007	Live rock 1
Cladophora	SV0008	Live rock 1
Cladophora	SV0009	Live rock 1
Caulerpa racemosa	SV0010	Live rock 1
Parvocaulis parvula	SV0011	Live rock 1
Caulerpa oligophylla	SV0012	Live rock 1
Palisada sp1	SV0013	Live rock 1
Chlorodesmis	SV0014	Live rock 1
Caulerpa racemosa	SV0015	Live rock 1
Boodlea sp1	SV0016	Live rock 1
Boodlea sp2	SV0017	Live rock 1
Sarconema filiforme	SV0019	Live rock 1
Caulerpa racemosa	SV0020	Live rock 1
Ulva sp1	SV0021	Live rock 1
Ulvella leptochaete	SV0022	Live rock 1
Caulerpa taxifolia	SV0023	Live rock 1
Gracilaria vieillardii	SV0024	Live rock 1
*Chaetomorpha vieillar*dii	SV0025	Live rock 1
*Chaetomorpha vieillar*dii	SV0026	Live rock 1
Chaetopmorpha sp3	SV0027	Live rock 1
Gracilaria vieillardii	SV0035	Live rock 1
*Pterocladiella caerules*cens	SV0036	Live rock 2
Pterocladiella sp1	SV0037	Live rock 2
Caulerpa cupressoides	SV0038	Live rock 2
Peyssonnelia sp5	SV0039	Live rock 2
Peyssonnelia sp3	SV0040	Live rock 2
Hydrolithon sp2	SV0041	Live rock 2
Hydrolithon sp3	SV0042	Live rock 2
Hydrolithon sp1	SV0043	Live rock 2
Peyssonnelia sp6	SV0044	Live rock 2
Peyssonnelia sp3	SV0046	Live rock 2
Peyssonnelia sp2	SV0047	Live rock 2
Valonia macrophysa	SV0048	Live rock 3
Plocamium sp1	SV0049	Public Aquarium 1
Crouania attenuata	SV0050	Public Aquarium 1
Womersleyella setacea	SV0051	Public Aquarium 1
Dictyota implexa	SV0052	Public Aquarium 1
*Rhodymenia ardissone*i	SV0053	Public Aquarium 1
Asparagopsis	SV0054	Public Aquarium 1
Ceramium codii	SV0055	Public Aquarium 1
Antithamnion	SV0056	Public Aquarium 1
Caulerpa prolifera	SV0057	Public Aquarium 1
Caulerpa taxifolia	SV0058	Public Aquarium 1
Caulerpa constricta	SV0059	Public Aquarium 1
Sargassum muticum	SV0060	Public Aquarium 2
Caulerpa prolifera	SV0061	Public Aquarium 2
Caulerpa sertularioides	SV0062	Public Aquarium 2
Caulerpa racemosa	SV0063	Public Aquarium 2
Halimeda disoidea	SV0064	Public Aquarium 2
Polysiphonia sp1	SV0065	Public Aquarium 2
Cladophora	SV0066	Public Aquarium 2
Cladophoropsis	SV0067	Public Aquarium 2
Peyssonnelia sp1	SV0068	Public Aquarium 2
Sporolithon sp1	SV0069	Public Aquarium 2
Mesophyllum sp1	SV0070	Public Aquarium 2
Halopteris filicina	SV0071	Public Aquarium 1

Taxon	Sample ID	Location
Cladophora	SV0072	Public Aquarium 1
Caulerpa constricta	SV0073	Public Aquarium 3
Codium	SV0074	Public Aquarium 3
Caulerpa taxifolia	SV0075	Public Aquarium 3
Yonagunia zollingeri	SV0077	Public Aquarium 3
Caulerpa serrulata	SV0078	Public Aquarium 3
Bryopsis sp3	SV0079	Public Aquarium 3
Bryopsis sp1	SV0080	Public Aquarium 3
Valonia utricularis	SV0081	Public Aquarium 3
Chaetomorpha	SV0082	Public Aquarium 3
Mesophyllum sp1	SV0083	Public Aquarium 3
Chaetomorpha vieillardii	SV0084	Public Aquarium 3
Cladophora	SV0085	Public Aquarium 3
Derbesia sp3	SV0086	Public Aquarium 3
Ceratodictyon repens	SV0087	Public Aquarium 3
Antithamnion	SV0088	Public Aquarium 3
Chaetomorpha	SV0089	Public Aquarium 3
Erythrotrichia carnosa	SV0090	Public Aquarium 3
Polysiphonia	SV0091	Public Aquarium 3
Cladophora albida/sericea	SV0092	Public Aquarium 3
Cladophora pellucida	SV0093	Public Aquarium 3
Derbesia sp4	SV0094	Public Aquarium 3
Cladophora pellucida	SV0095	Public Aquarium 3
Cryptonemia lomation	SV0096	Public Aquarium 3
Coelarthrum	SV0097	Public Aquarium 3
Derbesia	SV0098	Public Aquarium 3
Chondracanthus saundersii	SV0099	Public Aquarium 3
Sarconema sp1	SV0100	Public Aquarium 3
Botryocladia sp1	SV0101	Public Aquarium 3
Rhodymeniaceae	SV0102	Public Aquarium 3
Polystrata sp1	SV0103	Public Aquarium 3
Mesophyllum sp1	SV0104	Public Aquarium 3

Taxon	Sample ID	Location
Caulerpa parvifolia	SV0107	Private aquarium 1
Caulerpa sertularioides	SV0108	Private aquarium 1
Caulerpa chemnitzia	SV0109	Private aquarium 1
Halimeda minima	SV0110	Private aquarium 1
Caulerpa parvifolia	SV0112	Private aquarium 2
Botryocladia sp1	SV0113	Private aquarium 2
Acanthophora spicifera	SV0114	Private Aquarium 3
Acanthophora spicifera	SV0115	Private Aquarium 3
Hypnea valentiae	SV0116	Private Aquarium 3
Caulerpa parvifolia	SV0117	Private Aquarium 3
Caulerpa parvifolia	SV0118	Retail shop 1
Asparagopsis taxiformis	SV0120	Retail shop 1
Mesophyllum sp4	SV0122	Retail shop 1
Incendia sp1	SV0123	Retail shop 1
Botryocladia	SV0124	Retail shop 1
Bryopsis	SV0125	Retail shop 1
Halimeda minima	SV0126	Retail shop 1
Derbesia	SV0127	Retail shop 1
Halimeda gigas	SV0128	Retail shop 1
Ramicrusta sp1	SV0129	Retail shop 1
Polysiphonia sp1	SV0130	Retail shop 1
Meredithia sp1	SV0132	Retail shop 1
Caulerpa constricta	SV0133	Public Aquarium 3
Codium arenicola	SV0134	Public Aquarium 3
Caulerpa constricta	SV0136	Public Aquarium 4
Cladophora herpestica	SV0137	Public Aquarium 4
Yonagunia zollingeri	SV0138	Public Aquarium 4
Mesophyllum sp2	SV0139	Public Aquarium 4
Valonia utricularis	SV0140	Public Aquarium 4
Chaetomorpha vieillardii	SV0141	Public Aquarium 4
Yonagunia sp1	SV0143	Public Aquarium 4
Cladophora herpestica	SV0144	Public Aquarium 4

Taxon	Sample ID	Location
Derbesia sp1	SV0145	Public Aquarium 4
Ulva	SV0146	Public Aquarium 4
Polysiphonia	SV0147	Public Aquarium 4
Cladophora albida/sericea	SV0148	Public Aquarium 4
Bryopsis	SV0149	Public Aquarium 4
Cladophora albida/sericea	SV0150	Public Aquarium 4
Cladophora prolifera	SV0151	Public Aquarium 4
Cladophora prolifera	SV0152	Public Aquarium 4
Phymatolithon sp1	SV0153	Public Aquarium 4
Peyssonnelia sp4	SV0154	Public Aquarium 4
Ulva laetevirens	SV0155	Private Aquarium 4
Caulerpa constricta	SV0156	Private Aquarium 4
Laurencia sp1	SV0157	Private Aquarium 4
Chaetomorpha vieillardii	SV0158	Private Aquarium 4
Griffithsia sp1	SV0159	Private Aquarium 4
Hypnea sp1	SV0166	Private Aquarium 4
Haraldiophyllum sp1	SV0167	Private Aquarium 4
Halimeda opuntia	SV0169	Private Aquarium 4
Caulerpa cupressoides	SV0170	Private Aquarium 4
Halymenia durvillei2	SV0172	Private Aquarium 4
Halimeda gigas	SV0173	Retail shop 2
Halimeda gigas	SV0174	Retail shop 2
Dictyota ceylanica4	SV0175	Retail shop 2
Dictyota friabilis1	SV0176	Retail shop 2
Caulerpa racemosa	SV_0.1	Private Aquarium 5
Titanophora sp1	SV_0.10	Private Aquarium 5
Mesophyllum sp1	SV_0.11	Private Aquarium 5
Melobesioideae sp2	SV_0.12	Private Aquarium 5
Sporolithon sp1	SV_0.13	Private Aquarium 5
Melobesioideae sp2	SV_0.14	Private Aquarium 5
Titanophora sp1	SV_0.15	Private Aquarium 5
Sporolithon sp1	SV_0.16	Private Aquarium 5

Taxon	Sample ID	Location
Acrosymphyton sp1	SV_0.19	Private Aquarium 5
Botryocladia sp2	SV_0.2	Private Aquarium 5
Mesophyllum sp3	SV_0.20	Private Aquarium 5
Melobesioideae sp1	SV_0.3	Private Aquarium 5
Sporolithon sp3	SV_0.4	Private Aquarium 5
Mesophyllum sp1	SV_0.6	Private Aquarium 5
Sporolithon sp2	SV_0.7	Private Aquarium 5
Titanophora sp1	SV_0.8	Private Aquarium 5
Acrosymphyton sp1	SV_0.9	Private Aquarium 5
Asparagopsis taxiformis	SV_1.1	Live rock 4
Titanoderma sp1	SV_1.11	Live rock 4
Peyssonnelia japonica	SV_1.11A	Live rock 4
Lithophyllum sp4	SV_1.11B	Live rock 4
Lithophyllum sp1	SV_1.11C	Live rock 4
Pneophyllum	SV_1.12	Live rock 4
Porolithon	SV_1.13	Live rock 4
Neosiphonia sp1	SV_1.14	Live rock 4
Dictyota friabilis1	SV_1.16	Live rock 4
Bryopsis sp1	SV_1.17	Live rock 4
Ulva sp2	SV_1.19	Live rock 4
Caulerpa parvifolia	SV_1.2	Live rock 4
Lithophyllum sp2	SV_1.21	Live rock 4
Lithophyllum sp3	SV_1.21A	Live rock 4
Lithophyllum sp2	SV_1.21B	Live rock 4
Sporolithon sp3	SV_1.24	Live rock 4
Dictyota friabilis1	SV_1.26	Live rock 4
Ramicrusta sp1	SV_1.27	Live rock 4
Chaetomorpha vieillardii	SV_1.3	Live rock 4
Derbesia	SV_1.6	Live rock 4
Caulerpa flexilis	SV_1.7	Live rock 4
Dictyota friabilis1	SV_1.8	Live rock 4
Valonia macrophysa	SV_1.9	Live rock 4

Taxon	Sample ID	Location
Codium dwarkense	SV_2.1	Retail shop 3
Caulerpa parvifolia	SV_2.10A	Retail shop 3
Mastophoroideae sp1	SV_2.10BV	Retail shop 3
Valonia macrophysa	SV_2.10C	Retail shop 3
Derbesia sp3	SV_2.10D	Retail shop 3
Amphiroa	SV_2.11	Retail shop 3
Bryopsis sp3	SV_2.12	Retail shop 3
Ulvella	SV_2.13A	Retail shop 3
Peyssonnelia japonica	SV_2.13B	Retail shop 3
Haraldiophyllum sp1	SV_2.14A	Retail shop 3
Bryopsis sp2	SV_2.15	Retail shop 3
Caulerpa lentillifera	SV_2.16	Retail shop 3
Caulerpa parvifolia	SV_2.18	Retail shop 3
Harveylithon sp1	SV_2.19A	Retail shop 3
Harveylithon sp1	SV_2.19B	Retail shop 3
Cryptonemia sp1	SV_2.2	Retail shop 3
Sargassum sp1	SV_2.20	Retail shop 3
Haraldiophyllum sp1	SV_2.21	Retail shop 3
Haraldiophyllum sp1	SV_2.22	Retail shop 3
Caulerpa parvifolia	SV_2.23	Retail shop 3
Lithophyllum sp5	SV_2.28A	Retail shop 3
Mastophoroideae sp2	SV_2.28B	Retail shop 3
Ptilophora scalaramosa	SV_2.3	Retail shop 3
Halymenia durvillei1	SV_2.5	Retail shop 3
Caulerpa parvifolia	SV_2.6A	Retail shop 3
Cryptonemia sp1	SV_2.6B	Retail shop 3
Caulerpa taxifolia	SV_2.7	Retail shop 3
Peyssonnelia sp7	SV_2.9	Retail shop 3

Table S3. Species used for the thermal niche modelling analysis with their record count, midpoint of the thermal range, and the number of ecoregions they currently occur in and estimated under current (2010) and future (2055) climate conditions.

Species	Count	Midpoint (°C)	Current	2010	2055
Acanthophora spicifera	319	24.4	0	0	1
Asparagopsis taxiformis	545	21.5	7	8	10
Boergesenia forbesii	111	25.7	0	0	0
Caulerpa brachypus	127	22.3	0	4	6
Caulerpa chemnitzia	76	25.2	0	0	0
Caulerpa cupressoides	474	24.3	0	0	1
Caulerpa flexilis	245	17.1	0	2	2
Caulerpa lentillifera	181	24.5	0	0	1
Caulerpa prolifera	205	21.4	4	8	10
Caulerpa racemosa	954	23.1	1	3	5
Caulerpa serrulata	418	25.1	0	0	0
Caulerpa sertularioides	495	25	0	0	0
Caulerpa taxifolia	467	21.3	2	8	10
Ceramium codii	77	21.7	1	7	9
Cladophora albida	371	14.8	11	17	18
Cladophora herpestica	82	23.1	1	3	4
Cladophora pellucida	408	15.4	6	13	13
Cladophora prolifera	193	18.6	8	11	12
Cladophora sericea	920	12.1	6	8	8
Codium dwarkense	33	26.2	0	0	0
Crouania attenuata	111	18.6	6	11	12
Dictyota ceylanica	159	24.6	0	0	0
Dictyota friabilis	157	24.6	0	0	0
Dictyota implexa	52	19.1	5	11	12
Erythrotrichia carnea	537	15.6	9	16	18
Halimeda discoidea	639	24.7	0	0	0
Halimeda minima	161	25.2	0	0	0
Halimeda opuntia	739	25.4	0	0	0
Halopteris filicina	418	16.1	7	10	12
Halymenia durvillei	72	24.7	0	0	0
Hypnea valentiae	163	21.4	2	9	10
Parvocaulis parvulus	34	25.3	0	0	0
Pterocladiella caerulescens	71	24.3	0	0	0
Rhodymenia ardissonei	204	16	5	8	10
Sarconema filiforme	67	22.5	1	6	8
Sargassum muticum	1094	11.9	6	6	8
Valonia macrophysa	141	21.7	3	8	10
Valonia utricularis	105	19.9	6	10	10
Womersleyella setacea	94	19.7	7	7	10

Chapter 7

Modelling the past, present and future distribution of invasive seaweeds in Europe

Samuel Bosch[1,2], Eduardo Gomez Giron[1], Brezo Martínez[3] and Olivier De Clerck[1]

[1]*Research Group Phycology, Biology Department, Ghent University, Krijgslaan 281/S8, 9000 Ghent, Belgium*
[2]*Flanders Marine Institute (VLIZ), InnovOcean site, Wandelaarskaai 7, 8400 Ostend, Belgium*
[3]*Biology and Geology Department, Rey Juan Carlos University, Tulipán sn.,Móstoles 28933, Spain*

Manuscript in preparation.

Abstract

Invasive species can cause significant problems at ecosystem, economic and social levels. Assessing the potential geographic range of such species in invaded regions is therefore increasingly promoted for proactive ecological management. Unfortunately, because invasive species are by definition not at equilibrium within recipient environments, there is considerable uncertainty on how to model their distributions. In this study we evaluated the performance of species distribution modelling, trained with native and/or non-European distribution records, as a tool for predicting the spread of invasive seaweeds at various stages of the invasion process. We estimated the level of niche expansion observed under analog and non-analog conditions and assessed which areas in Europe are expected to be disproportionally impacted by migrations of introduced seaweeds due to climate change. Our results indicate that due to considerable niche expansion in non-analog conditions including only native records is generally not sufficient to predict the range of invasive species. Including distribution records from non-European invaded regions on the other hand significantly increases the predictive power of the models and reduces the measured niche expansion in analog and non-analog conditions considerably. The European change and turnover maps combined with an assessment of the uncertainty therein predict an increased habitat suitability in northern Europe (northern UK, Scandinavia, Iceland), while southern European are likely to become less suitable. In addition to the overall picture, uncertainty in the estimates is apparent for specific regions but correlates only moderately to changes in habitat suitability.

Introduction

Invasive species rank as one of the greatest threats to marine coastal biodiversity (McGeoch et al., 2010; Seebens et al., 2013). The European shores sadly stand out as a hotspot for introduced species (Molnar et al., 2008) and seaweeds represent one of the largest groups of marine aliens, accounting for 10 to 30% of all marine introduced species in Europe (Schaffelke et al., 2006; Williams & Smith, 2007; Zenetos et al., 2012; Katsanevakis et al., 2013). Several seaweeds, such as kelp or fucoids in the Atlantic Ocean and the canopy-forming *Cystoseira* species in the Mediterranean Sea, are true ecosystem engineers or foundation species (Jones et al., 1994; Mineur et al., 2015). Consequently, changes in seaweed communities can provoke cascading effects influencing the entire ecosystem, including for example changes in abundances of herbivores and understorey coralline algae (Monteiro et al., 2009; Harley et al., 2012; Verges et al., 2014; Wernberg et al., 2016).

The ecological importance of seaweeds combined with high introduction rates of non-native species, highlights the need for methods able to accurately predict the future distribution of invasive species, preferably at the early stages of the invasion process. Species distribution modelling (SDM) links species occurrences with the environmental characteristics, and has the potential to predict distributions in a geographically explicit framework, including extrapolation in space and time. SDM can be used to identify areas with suitable habitat, assess whether introductions are likely to be successful, anticipate arrival points, and predict the extent of potential spread following an introduction. However, arrival points also heavily depend on the introduction vector (e.g. shipping, aquaculture) and the level of human activity related to these introduction vectors (Reiss et al., 2015). SDM can thus, supplemented with information on introduction vectors, help us inform decisions about preventive and control actions.

The predictive power of SDM, however, is very much dependent on the assumption that species are at equilibrium with their environment, which implies that distribution records reflect stable relationships with environment. The very nature of invasive alien species, which are possibly still in the process of range expansion in the introduced range, means that this assumption is not met for these organisms (Elith et al., 2010). Furthermore, the biotic interactions in the native and introduced environment may differ leading to changes in the geographical and environmental range (DeWalt et al., 2004; Mitchell et al., 2006). Therefore SDM of invasive or range-shifting species is particularly challenging, and requires the development of advanced modelling techniques potentially integrating mechanistic and correlative approaches (Kearney & Porter, 2009). Mechanistic approaches may model species

distributions by modelling the body temperature based on functional traits of organisms instead of using the air or sea surface temperature as indicators of environmental stress (Kearney et al., 2010; Helmuth et al., 2011) However, data availability for mechanistic models is limited while distribution data to build correlative models is more widely available (Elith et al., 2010). Improving transferability of correlative models to other time/space datasets has been accomplished by reducing overfitting and sample selection bias . Such models can be obtained by using different model choices in the background selection (Barbet-Massin et al., 2012; Martínez et al., 2015), restricting model complexity (regularization and number of variables) (Wenger & Olden, 2012), eliminating sample selection bias (Verbruggen et al., 2013; Radosavljevic & Anderson, 2014) or applying ensemble models (Hijmans & Graham, 2006; Araújo & New, 2007).

In order to accurately predict the introduced geographic range, the environmental niche of the native and introduced populations of the species should be similar (Guisan et al., 2014). While, Wasof et al. (2015) have shown that environmental niches are generally conserved between separated populations of alpine plants, for introduced seaweeds this has not been shown. We, furthermore, distinguish niche expansion or niche conservatism in analog and non-analog conditions (Guisan et al., 2014). Analog conditions are environmental conditions occurring both in the native and invaded range, while non-analog conditions are only occurring in one of the ranges. Although calculating niche change metrics in non-analog climates provides little insight in the evolution of the niche of a species, the change in niche metrics in non-analog conditions is still highly relevant for predicting the distribution of the species in invaded ranges (Petitpierre et al., 2012; Webber et al., 2012; Guisan et al., 2014).

In this study, we aim to improve predictions of invasive seaweeds in Europe and to map areas that will be disproportionally affected by changes in invasive seaweed distributions due to climate change. To this end, we first analyse the predictive performance of SDM towards the identification of suitable habitats in Europe at different stages of the invasion history based on a case study with five invasive seaweeds. Next, we calculated niche expansion and relate it to the performance of SDM. Finally, using an expanded dataset of 15 commonly recorded and widespread non-native seaweeds, we identified geographic risk areas in Europe by comparing current and future climate species distribution models.

Methods

Records collection

In order to explore the modelling of invasive seaweeds and their environmental niche in Europe we collected species records for five invasive species, for which we have ample distribution records in the introduced and native range: *Codium fragile* subsp*. fragile, Dictyota cyanoloma, Grateloupia turuturu, Sargassum muticum* and *Undaria pinnatifida*. Distribution records were classified as native or invasive, by region (Asia, Europe, America, Africa and Australia) and by year whenever possible. With respect to *C. fragile* subsp. *fragile* we decided to include species records for all subspecies as the identification of subspecies of *C. fragile* is notoriously difficult based on morphological criteria and the invasive subspecies is found in the entire range of the species (Brodie et al., 2007b; Provan et al., 2008; McDonald et al., 2015). Moreover, DNA barcodes and morphometric data indicate that *C. fragile* may actually consist of two species, the invasive subspecies *fragile* and a second species grouping all remaining subspecies (Verbruggen et al., 2016). Distribution records were collected from different data portals including the Macroalgal Herbarium Portal (macroalgae.org), Global Biodiversity Information Facility (gbif.org), Australia's Virtual Herbarium (avh.chah.org.au), Natural History Museum London (nhm.ac.uk), Muséum National d'Histoire Naturelle (mnhn.fr) with records updated until March 2016. For the last part of our study we want to uncover areas in Europe that will be affected by displacements of introduced seaweeds due to climate change. Therefore, we collected occurrences for an additional set of ten seaweeds: *Asparagopsis armata*, *Bonnemaisonia hamifera*, *Colpomenia peregrina*, *Dasya sessilis*, *Dasysiphonia japonica*, *Gracilaria vermiculophylla*, *Grateloupia subpectinata*, *Lomentaria hakodatensis*, *Polysiphonia harveyi* and *Polysiphonia morrowii*. Together with the five species from the first part they form a set of 15 representative and widely introduced seaweeds in Europe.

The quality of the distribution records was checked by geographic visualization and verification of mismatches between the location where the records were found and the coordinates recorded (Marcelino & Verbruggen, 2015). Duplicate records were eliminated as well as records located in the same grid cell of the environmental data in the same year. Records within the boundaries of the landmask were moved to the nearest ocean grid cell if located within 1,000 meters from an ocean grid cell. Records further than 1,000 meters from an ocean cell were deleted.

Environmental data

The Bio-ORACLE dataset was used as a source for environmental predictor variables. It consists of global rasters with a spatial resolution of 5 arcmin (Tyberghein et al., 2012) and it is primarily designed for global-scale niche modelling of shallow water marine organisms (Marcelino & Verbruggen, 2015). The environmental layers were retrieved using the *sdmpredictors* R package (Bosch et al., 2016).

Predictor selection is a major concern when building transferable SDMs. Many studies have addressed the consequences of variable selection (Rödder & Lötters, 2009; Verbruggen et al., 2013; Barbet-Massin & Jetz, 2014). The problem underlying this issue is the absence of causal links between predictor and response variables which may constrain the predictive power of the model (Austin, 2002; Martínez et al., 2015). In this study, the selection of variables was made a priori, taking general knowledge on the physiology and ecology of seaweeds into account (Lüning, 1990; Hurd et al., 2014). In addition variables with high correlation were not selected.

Four variables were selected a priori as potentially influencing seaweed distributions (Table 1). Sea surface temperature is suspected to be the main variable driving the distribution of seaweeds. It can affect the performance of growth, photosynthesis, reproduction and survival (Breeman, 1988; Lüning, 1990; Eggert, 2012). We used two temperature measures: maximum sea surface temperature and sea surface temperature range. Two more variables were added: mean photosynthetically active radiation and sea surface salinity. Seaweeds are photosynthetic organisms and therefore the quantity of light can affect their growth and limit habitat suitability. Salinity can influence osmotic dynamics limiting nutrient absorption and affect membrane integrity (Hurd et al., 2014), thus influencing growth, fitness and survival of seaweeds and therefore limit suitable habitats (Martins et al., 1999), as for example is the case in the Baltic Sea (Nyström Sandman et al., 2013).

Table 1. Overview of the ranges (minimum, median and maximum) of the environmental data used for modelling invasive seaweeds both for global and coastal data with the values for Europe between brackets. The different layers are maximum and range of sea surface temperature (SST), mean photosynthetically active radiation (PAR) and sea surface salinity.

	Minimum		Median		Maximum	
Layer	Global	Coastal	Global	Coastal	Global	Coastal
SST (max)	-1.5 (0.7)	-1.5 (0.6)	25.3 (19.2)	20.2 (19.9)	35.9 (32.7)	37.6 (34.4)
SST (range)	0.1 (2.5)	0.0 (1.3)	4.1 (7.0)	5.3 (12.2)	29.6 (25.1)	31.2 (29.4)
PAR (mean)	0.5 (23.7)	0.5 (8.2)	39.6 (31.3)	34.4 (32.6)	52.3 (47.1)	66.9 (55.0)
Salinity	0.0 (2.1)	0.0 (1.8)	34.7 (35.5)	33.4 (33.9)	40.7 (40.6)	41.5 (41.5)

Distribution modelling

Species distributions were modelled using four different algorithms: surface range envelope, which is equivalent to bioclim (SRE, Busby, 1991), generalized linear model (GLM), maximum entropy (MaxEnt, Phillips et al., 2004) and random forests (Breiman, 2001). For MaxEnt and GLM, we explored the complexity of the models fitted by building models with linear and quadratic features. The complexity of SRE cannot be controlled and for random forests different settings were not explored. Additionally, an ensemble model (Araújo & New, 2007) was built by averaging the results of the most transferable models. Distributions were modelled using the R (R Core Team, 2016) packages *biomod2* (Georges & Thuiller, 2013; Thuiller et al., 2016) and *dismo* (Hijmans et al., 2016).

Sample selection bias is one of the main problems impacting the transferability of models, leading to an overrepresentation of conditions in places where collecting effort is higher and thereby inflating model performance indices (Hijmans, 2012). In order to reduce sample selection bias, and therefore also environmental bias, a spatial occurrence thinning method was used (Veloz, 2009). Presence records were eliminated with the R package *spThin* (Aiello-Lammens et al., 2015) for two different thinning distances, 30 and 100 kilometres, and the results of these were compared for the different model algorithms and complexities.

Presence-only methods use species occurrence records and background points, which are selected randomly in the study area. As suggested by Phillips & Dudík (2008) we used 10000 background points in our study, which is adequate to cover the whole area. Two different sets of training background points were generated, one with points restricted to all pixels adjacent to land (coastal) and one with global background points.

In order to evaluate the different modelling options an evaluation dataset is needed (Arlot & Celisse, 2010). As no independently sampled evaluation data was available, three different ways to obtain the evaluation dataset from our dataset were explored by either splitting 'randomly', 'temporally' or 'spatially' (Roberts et al., 2016). The random splitting method consists of randomly splitting the dataset into training and testing. In the temporal approach, we used records from the earlier years to build the model and more recent occurrence records to evaluate them. Testing absence points were selected using pairwise distances such that the distance between the test occurrences and pseudo-absence points is the same as the distance between training and test occurrences (Hijmans, 2012). Finally, the spatial approach consists of dividing datasets based on geography. European

occurrence records and coastal pseudo-absence points are used to evaluate the model built with records outside of Europe. The European region was determined as the area with longitude between -34° and 65° and latitude between 29° and 73°.

Three different metrics were used to evaluate model performance: area under the receiver operating curve (AUC) (Hanley & McNeil, 1982), Cohen's kappa and the H-measure. Although AUC values have been criticized in the context of species distribution modelling (Lobo et al., 2008), its use was motivated because it is objective, threshold independent and insensitive to imbalanced datasets (Hand, 2009). Cohen´s kappa measures the agreement between predictions of the model and observations but corrects for agreement expected by chance. Kappa is sensitive to imbalanced datasets. Therefore, it has been corrected by creating a multitude of kappa values calculated from random balanced subsamples and taking the first quartile as the final kappa value. The H-measure (Hand, 2009), is similar to AUC but has as additional property that it is independent of the distribution of the empirical scores.

Starting from the model choices resulting in the most transferable and robust models, SDMs were created for all five species for different timeframes in their invasive history. Models were fitted with an increasing number of records starting with all records from the last year prior to introduction in Europe (T1). We assessed the ability of these and successive models (T2, T3 and T4), which cumulatively included more invasive records, to predict the European distribution.

This invasive history was analysed for two scenarios: a restricted and a global one. The restricted scenario consisted of a first model (T1), built with all native records and subsequent models with invasive records from Europe cumulatively added according to the invasive history (Table 2). The global scenario, on the other hand, consisted of models fitted with all available native and invasive records at the specific timeframe. Therefore, T1 models included native and invasive records from all non-European areas known at T1.

As continuous model projections are sometimes difficult to interpret, the creation of binary presence/absence maps can be a useful tool for risk assessment. We used as threshold the value that maximizes the sum of sensitivity and specificity (maxSSS) to create binary suitability maps because it is one of the best performing methods to create threshold maps when absences are not reliable (Liu et al., 2013).

Niche shifts

In order to measure the overlap in the realized niche of the species between the native and invaded range, three different indices have been calculated: niche expansion, niche stability and niche unfilling (Guisan et al., 2014). Niche expansion measures conditions in niche space not occupied in the native range. On the other hand, niche stability measures the conditions shared between both distributions. The niche stability is comparable to the niche overlap as assessed through Schoener's D or Hellinger's I. Lastly, niche unfilling measures the conditions occupied in the native range but not in the invasive range (Guisan et al., 2014). Ordination techniques, more specifically PCA-env, have been show to measure the niche overlap between two distributions better than SDM methods (Broennimann et al., 2012). The PCA-env method compares kernel smoothed species occurrence densities in an ordinated environmental space, which allows for direct comparisons of species-environment relationships in environmental space. Similar to the distribution modelling, niche measures were calculated for the restricted and global scenario and by either only taking into account analog conditions or also including non-analog conditions (Guisan et al., 2012; Webber et al., 2012). The study area used to distinguish between analog and non-analog conditions was defined as all ecoregions (*sensu* Spalding et al., 2007) wherein occurrences are located. All indices were transformed to percentages of the entire niche. These analyses were performed using the R package *ecospat* (Broennimann et al., 2016).

Risk areas

In order to identify areas at risk in Europe we modelled current and future climate distributions with the same predictors used to build transferable models. Additionally, we fitted models for two extra predictor sets by substituting maximum sea surface temperature with the mean or minimum sea surface temperature, since according to Synes and Osborn (2011) it is rarely clear which temperature variable is most applicable.

Ensemble models for these three sets of predictors were created by averaging the output of SDMs built with coastal background and all occurrence records using generalized linear models (GLM) with quadratic features, MaxEnt with quadratic features, random forests (RF) and surface range envelope model (SRE). These current climate models were subsequently projected to the three IPCC climate scenarios B1 (550 ppm stabilization), A1B (720 ppm stabilization) and A2 (>800 ppm) for the year 2100 (Jueterbock et al., 2013). The maximum sum of sensitivity and specificity (maxSSS) was used as a threshold for converting the current and future climate SDMs to binary maps. These binary maps for the 15 species were summed to

get a map of the number of species predicted in the current and future climate. The change maps were obtained by subtracting the current and future climate count maps for the three predictor sets, and subsequently calculating the mean and standard deviation of these. The mean anomaly map then reflects the change in number of species in the different areas, and the standard deviation map indicates the uncertainty of the results. Additionally, maps of the mean and standard deviation of the species turnover were calculated by counting, based on the binary maps, for each raster cell the number of species that either are predicted in the current climate and are not predicted in the future climate or vice versa.

Results

Distribution modelling

The most transferable species distribution models were obtained by creating an ensemble based on MaxEnt and GLM with quadratic features, random forests and SRE, with models fitted using coastal background and species specific spatial thinning settings. Further information on the modelling choices is provided in the Supporting information.

Based on these modelling choices we present the model performance when cumulatively more records are included in the training set representing different time points during the history of the introduction process. We repeated this process for two different setups, the first one using all non-European records (global scenario) and the second one using only native records (restricted scenario). Fig. 1 represents the performance of the model for the two different scenarios for all species, while the previously introduced Table 2 gives an overview of the number of records available for each timeframe.

For the restricted scenario, when only native records are used to build the model, AUC values are generally lower (left most values in Fig. 1A). The highest AUC is measured for *D. cyanoloma* (AUC = 0.872 with 11 records from the native range). While models for *U. pinnatifida* (AUC = 0.618) and *S. muticum* (AUC = 0.607) perform similarly with 77 and 71 records in the native range, respectively. Both *G. turuturu* (AUC = 0.536 with 43 records) and *C. fragile* (AUC = 0.485 with 44 records) have the lowest AUC. The AUC increases when occurrence records from Europe are included in the model (second, third and fourth value for each line in Fig. 1A and B), and this increase in AUC is larger at the early compared to the later phases of the introduction process.

Table 2 Data used to build the models in the global and restricted scenarios. The number of records included in the models (T1, T2, T3 and T4) is determined by the cumulative sum of the records (Timeframe). The number of records for the restricted scenario is determined by the sum of the Native and European records and for the global scenario by the sum of the Native, European and Non-European records. All records, including records after the last timeframe or without a year indication, were only used for calculating the niche metrics.

Species	Timeframe	Native	European	Non-European
Codium fragile	Before 1845 = T_1	44	0	9
	Before 1940 = T_2		41	49
	Before 1965 = T_3		169	164
	Before 1990 = T_4		471	350
	All		965	917
Dictyota cyanoloma	Before 1935 = T_1	11	0	0
	Before 2008 = T_2		5	0
	Before 2010 = T_3		15	2
	All		41	2
Grateloupia turuturu	Before 1969 = T_1	43	0	1
	Before 1985 = T_2		13	1
	Before 2000 = T_3		44	3
	All		170	36
Undaria pinnatifida	Before 1971 = T_1	77	0	0
	Before 1990 = T_2		7	1
	Before 2000 = T_3		50	4
	All		165	51
Sargassum muticum	Before 1970 = T_1	71	0	93
	Before 1975 = T_2		12	112
	Before 1985 = T_3		143	159
	Before 2000 = T_4		412	179
	All		1447	345

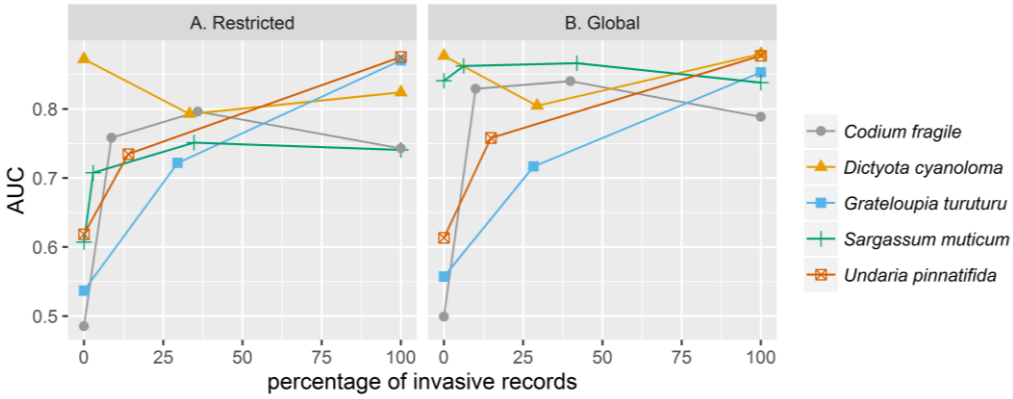


Figure 1 Evolution of the area under the curve (AUC) values at different points along the invasive history. The left figure (A) shows the AUC values for the restricted scenario, in which at T1 only native records are used for model fitting. Subsequent time points use both native and European records. For the right figure (B) all occurrence records (European and non-European) known at the specific year are used for modelling. The x-axis represents the percentage of invasive records used to build the model with the total number of records included in the last model as 100%.

Model performances of the global scenario are quite similar to those of the restricted scenario with the exceptions for *C. fragile* and *S. muticum* (Fig. 1B). Both the models for *S. muticum* and *C. fragile* have markedly higher AUC values when, next to native records, invasive records from other parts of the world known before the introduction in Europe (T1) are used to create the SDM. The other evaluation metrics (H-measure and kappa) show similar trends (Fig. S2 in Supporting information).

Fig. 2 shows maps of the model predictions of *S. muticum* for the timeframes T1 and T4 for the restricted and global scenarios. While both T1 models generally predict low habitat suitability, the threshold map of the global scenario overall reflects the present European invaded area well (Fig. 2B). However, the model failed to predict parts of the French and Catalonian coasts in the Mediterranean Sea. Model predictions from the restricted scenario (Fig. 2A and C) tend to overpredict the Mediterranean and Baltic sea and underpredict Portugal and the North of the British Isles. Maps of the other species are available in Figs. S3 to S6 in Supporting information.

Niche shifts

The niche analysis was performed in both the restricted and global scenario and niche indices were measured with or without taking into account non-analog conditions. Generally very low niche unfilling was measured with the highest value being 3% for *G. turuturu*. From Fig. 3, which reports the niche expansion, we notice that except for *D. cyanoloma* there is virtually no niche expansion between the invasive records in Europe and the non-European records, regardless of whether non-analog conditions are included in the niche expansion. When the niche of the native records is compared with the niche of the European records (restricted scenario), we see that in analog conditions there is 20 % niche expansion for *C. fragile* and about 10 % niche expansion for *S. muticum*. Niche expansion is highest when non-analog conditions are also taken into account with almost 50% for *C. fragile*, around 20% for *G. turuturu* and *S. muticum* and 10 % for *U. pinnatifida*. *D. cyanoloma*, which is not introduced in regions outside Europe, has less than 10 % niche expansion between native and invaded range when non-analog conditions are taken into account.

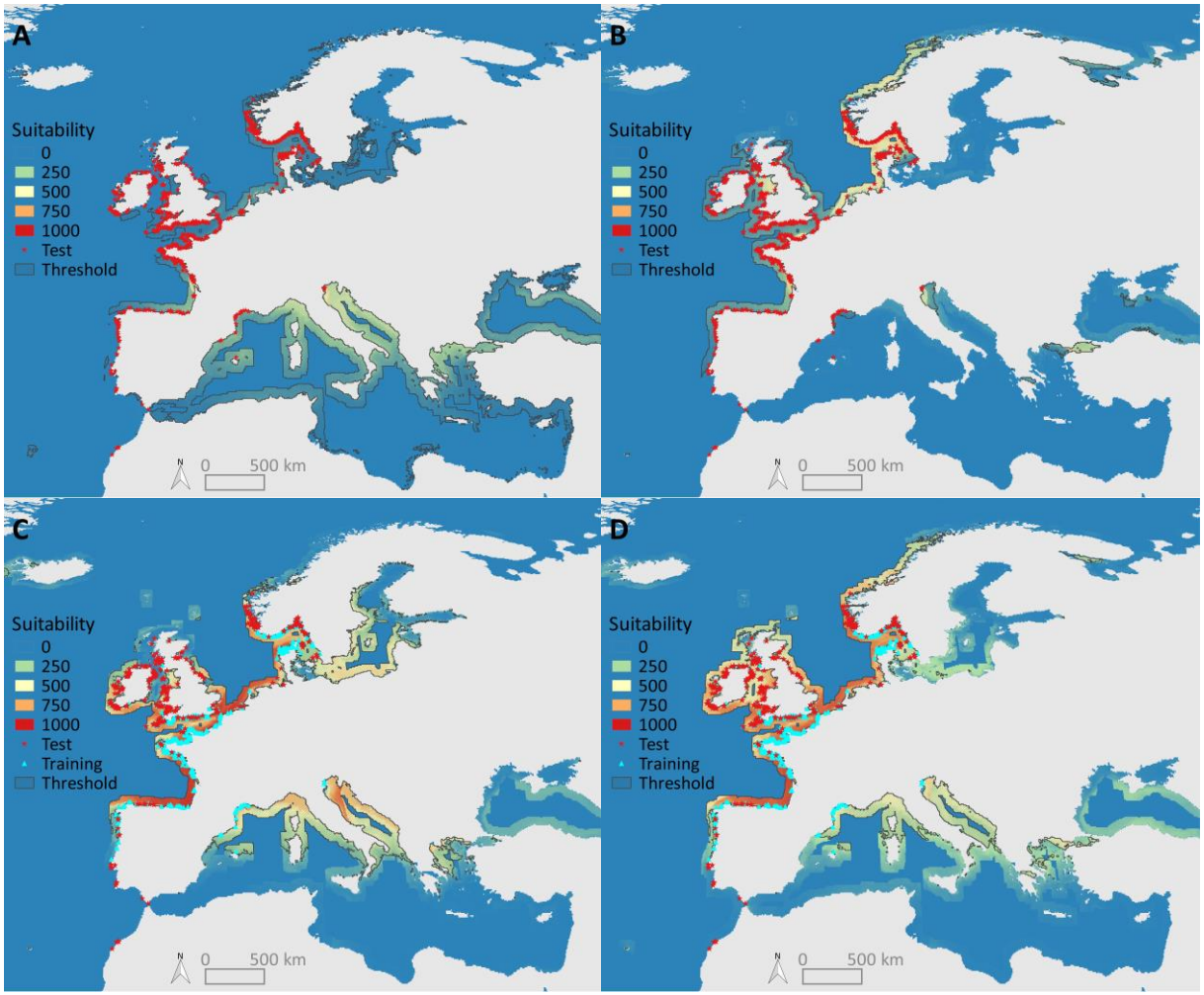


Figure 2. European predictions of suitable areas for *Sargassum muticum*. Red stars are locations used as test occurrences and the training records are in cyan triangles. The records for fitting the models are: A) only native records, B) native records and all invasive records known before the introduction of *S. muticum* in Europe, C) native records and all European records known in the year 2000 (T4) and D) native records and all invasive records known in 2000. A) and C) represent models from the restricted scenario while B) and D) are models from the global scenario. For further information about the number of records included in each model we refer to Table 2. Maps of the other species are available in Supporting information (Figs S3-S5).

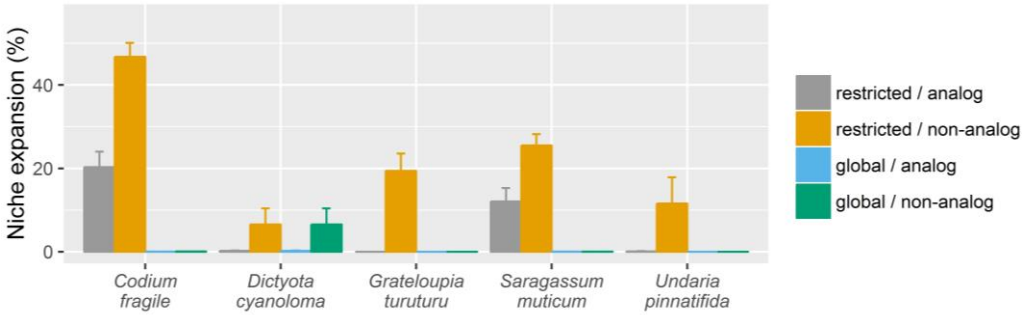


Figure 3 Niche expansion between the native and European occurrence records for four different setups. The setups have differences in the records used for comparing niches and in what is considered as niche expansion. We compared the niche expansion in Europe for the native records (restricted) and the native and non-European introduced distribution records (global). For the analog scenario niche expansion is only measured in environmental space that is available in both native and invaded area. In the non-analog scenario all niche expansion is reported. The error bars represent the standard error of using either 30 or 100 km spatial thinning.

Risk areas

Regarding the assessment of areas at risk in Europe, we see that the largest increase in number of introduced species is predicted in the northern areas of Europe, more specifically along the coasts of Iceland, Denmark and Norway by 2100 for the IPCC scenario B1 (Fig. 4A). Smaller increases in the number of introduced species are predicted for the United Kingdom, the Netherlands and Belgium. Areas with the biggest decreases, effectively becoming less suitable for the modelled list of species, are mostly located in the Mediterranean region. Additionally, some smaller spots in the Atlantic show a decrease in the number of introduced species predicted. The standard deviation map (Fig. 4B) clearly shows that some areas with larger gain also have a higher uncertainty and that the northern regions have higher standard deviations. The Pearson correlation between the absolute value of the mean and standard deviation of the change maps is only 0.55 (Table 4), which indicates a low to moderate correlation. The turnover maps for the same IPCC scenario B1 (Fig. 4C) reveal a high turnover for the regions with big increases and decreases as identified by the change maps and a number of additional areas with changes in species composition. This is most notably the case for the southern coasts of Great-Britain and Ireland. The Pearson correlation between the mean and standard deviation of the turnover maps is 0.66 (Table 4).

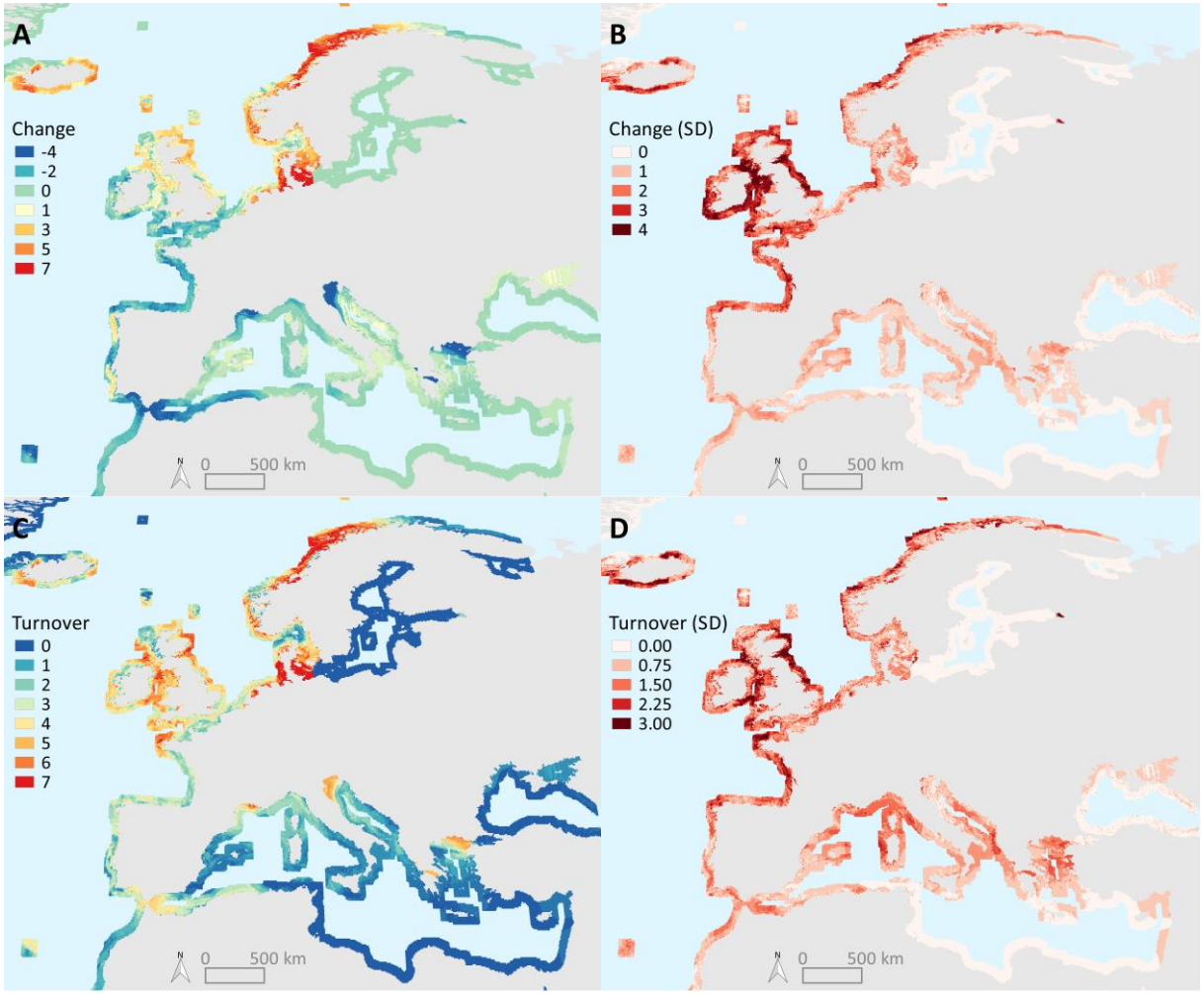


Figure 4. The change in number and turnover of introduced seaweeds predicted by 2100 under the IPCC climate change scenario B1. The top left map (A) shows the mean difference in number of introduced species between the current climate and climate change scenario B1 from SDMs build for 15 species. The top right map (B) indicates the standard deviation in model predictions when using the mean, minimum or maximum sea surface temperature as one of the four predictors for building the distribution models for each species. The bottom left (C) and right (D) maps indicate the mean and standard deviation of the turnover of introduced seaweeds. For the results based on two other IPCC scenarios, A1B (Fig. S7) and A2 (Fig. S8) we refer to Supporting information.

For the two other IPCC scenarios, A1B (Fig. S7) and A2 (Fig. S8) in Supporting information, we observe the same trends. This is confirmed by the Pearson correlation between the change and turnover maps of the different climate change scenarios which are all highly or very highly correlated, with the smallest correlation for the mean maps being 0.88 and for the standard deviation maps 0.82 (Table 4).

Table 4 Pearson correlation between the mean and standard deviation (SD) of the change (left) and turnover (right) maps for the different climate change scenarios (B1, A1B and A2). The left numbers are the correlations between the change maps, the right numbers represent the correlations between the turnover maps.

		B1		**A1B**		**A2**	
		Mean	SD	Mean	SD	Mean	SD
B1	Mean	1 / 1					
	SD	0.55 / 0.67	1 / 1				
A1B	Mean	0.90 /0.93	-	1 / 1			
	SD	-	0.85 / 0.82	0.53 / 0.66	1 / 1		
A2	Mean	0.88 / 0.91	-	0.95 / 0.97	-	1/ 1	
	SD	-	0.82 / 0.79	-	0.88 / 0.84	0.50 / 0.69	1 / 1

Discussion

Distribution modelling

Modelling the distribution of invasive species requires extrapolation to locations where the species have not previously been recorded. Therefore, general models with high transferability are needed (Randin et al., 2006). Several studies have researched methods to split occurrences into evaluation and training sets (Arlot & Celisse, 2010; Hijmans, 2012; Radosavljevic & Anderson, 2014; Roberts et al., 2016). The results of this study (Table S1 and Fig. S1) show that the random splitting method inflates the values of the evaluation metrics due to two different causes. Firstly, the environmental space distribution of the training and test occurrence records is the same. Hijmans (2012) already stated that closer testing and training points lead to artefacts in model evaluation and inflation of evaluation metrics. Secondly, background points and occurrences are more distant from each other in the environmental space leading to higher performance metrics. On the other hand, the temporal splitting approach shows low performance because background points and occurrences are distributed in the same environmental space rendering the differentiation between background points and presences difficult. With the spatial splitting approach, background points are well discriminated from occurrences but the occurrences used for testing the models and those used to build the model have a different distribution. In this context, we conclude that the spatial approach is the

most realistic cross-validation method for model selection of invasive species distribution models. This approach corroborates results by Radosavljevic & Anderson (2013), who demonstrated the power of geographic approaches to split the data to improve transferability and, therefore, to model invasive species.

Models with overly complex response curves limit the transferability of SDMs due to overfitting (Wenger & Olden, 2012; Merow et al., 2013; Verbruggen et al., 2013; Duque-Lazo et al., 2016). However, our results showed that including only very simple features results in models with a low performance as compared to using quadratic features, indicating that using only linear features results in underfitting (Hastie et al., 2009; Merow et al., 2014; Moreno-Amat et al., 2015). This can potentially be explained by the inability of the models to capture the relationship of predictors like maximum sea surface temperature (SST (max)) with the species distribution as an organism's response to temperature behaves like a quadratic curve and not a linear curve. With respect to algorithms we used GLM, MaxEnt, RF and SRE. GLM and MaxEnt generally performed well when models were tested with an independent European dataset (Table S3). However, a general trend does not exist which could be due to the variability in model performance of algorithms for different species (Elith et al., 2006; Araújo & New, 2007).In this context ensemble models prove to be a good solution to capture differences between model algorithms in a single transferable SDM (Table S3).

Background selection has a big impact on model transferability. The model performance of the coastal background is consistently higher (Table S3). The motivation for this approach was the impossibility of seaweeds, being coastal organisms, to survive in deep oceanic areas (Lüning, 1990; Marcelino & Verbruggen, 2015) and the usage of a similar approach in previous studies (Pauly et al., 2011; Martínez et al., 2015). Masking out training background data from the middle of the ocean improved model transferability when evaluated with testing records and coastal background points from Europe. Disregarding the absence of suitable substrate, the open ocean could hold environmental conditions suitable for the species. But when they are included in the background data they have to be classified as absences as species are not able to live there. By removing open ocean background data from the training set the number of false absences is thus significantly decreased, resulting in better and more transferable SDMs. Interestingly, differences between occurrence thinning parameters were species dependent. The lack of a common trend results from the idiosyncratic nature of the invasive process and recording history of the individual species. We followed the recommendations made in the literature (Phillips et al., 2009; Anderson & Raza,

2010; Barnes et al., 2014) and selected the wider thinning distances for those species with a small difference in performance for the two thinning distances since sample selection bias is one of the main drivers constraining transferability.

Threshold maps of the predictions of the European invasion allow us to visualize areas for which presences and absences were incorrectly predicted (Fig. 2). Species distributions were not predicted accurately when models were built without any invasive records from Europe. However, prediction improves rapidly when only 10 per cent of the European records were included to build the model. The addition of distribution records from other regions (global scenario case) improves the prediction and model performance mainly in *C. fragile* and *S. muticum* as these are the species with more records from other invaded regions. Especially, for *S. muticum* this resulted in a marked improvement of the T1 and later models. This implies that the inclusion of, the mostly Californian, invasive records known before the invasion in Europe add essential information about the environmental niche of *S. muticum* for modelling the European distribution. The fact that models perform generally better when including more records and when records from other invaded areas are included supports the idea that the whole environmental niche may not be recorded in the native range. Our results confirm that correlative models which aim to predict biological invasions should use all available records in order to capture the environmental niche better (Broennimann & Guisan, 2008; Verbruggen et al., 2013). But, in order to further improve the performance of species distribution models, next to extensive sampling of the native area, additional factors such as eco-physiological data and biotic interactions may need to be included.

For the other species the performance of T1 models is really low even if they are built with a relatively high number of records. For example the T1 model of *U. pinnatifida* was built with 77 native records and was barely able to predict the records in Europe. *D. cyanoloma* is a special case due to the few number of records available. The low number of records in *D. cyanoloma* models could explain the high AUC values and, therefore, it is probably heavily affected by the stochasticity of the known invasion process. Other factors contributing to the low initial performance include: differences in distributions in the environmental space between native and European occurrences, oversampling of specific native areas, lack of knowledge of the native distribution and lower competition in the invaded area.

Niche shifts

When using only analog climatic conditions, as suggested by Petitpierre et al. (2012), *C. fragile* and *S. muticum* are the species with the highest niche expansion in the

restricted scenario with both more than 10%, considered by Strubbe et al. (2013) to be a significant amount of niche expansion. The inclusion of distribution records from other invaded areas reduced the niche expansion to nearly zero. These results are very similar to previous studies of non-native plants and birds, with respectively 7 out of 50 and 8 out of 28 species displaying more than 10% niche expansion in analog conditions (Petitpierre et al., 2012; Strubbe et al., 2013). In contrast to non-native plants and birds, where niche unfilling was more prevalent than niche expansion, no niche unfilling was measured. This might potentially indicate a lack of sampling in the native range.

However, we agree with Webber et al. (2012) that studies aiming to forecast biological invasions should include non-analog conditions, as those studies are based on extrapolation in analog but also non-analog conditions. The difference of niche expansion with or without inclusion of non-analog conditions in the restricted scenario is 20% for *C. fragile* and *G. turuturu* and more than 10% for *S. muticum* and *U. pinnatifida* which could significantly constrain the prediction of introduced species. However the inclusion of records from invaded areas outside of Europe eliminated all significant niche expansion. This might explain why including other invaded records resulted in an improvement of the distribution models.

Risk areas

The increase in number of introduced species in the more northern areas of Europe is in accordance with Jueterbock et al. (2013) who predicted a northward shift for three North Atlantic seaweeds. But, the predicted risk areas are influenced by the fact that we only took into account previously known invasive seaweeds in Europe, for instance the predicted decrease in introduced species in the Mediterranean Sea was to be expected given that the rising temperatures in the Mediterranean will render it unsuitable for several of the modelled species. This doesn't necessarily imply that new species will not be introduced as the increased temperature might make it suitable for other, predominantly subtropical to tropical species that have not yet been reported in Europe. The big increase in predicted suitability for invasive seaweeds in the Northern Atlantic by 2100 might be tempered by the fact that the introduction and distribution of species depends on more factors than only the environmental suitability. Although temperature can be considered to be the main factor restricting the distribution of seaweeds (Breeman, 1988; Lüning, 1990; Eggert, 2012), it is also limited by other abiotic (bathymetry, light, substrate) and biotic factors (competition and grazing) at smaller geographic scales (Marcelino & Verbruggen, 2015).

Another inherent factor of uncertainty in the results is the fact that we used models to predict the distributions of species in the current and future climate. One of the important factors contributing to this uncertainty is the selection of predictor variables (Synes & Osborne, 2011). By calculating the mean and standard deviation of models built with the minimum, mean and maximum temperature we tried to mitigate and visualize this uncertainty. By employing spatial thinning we reduced sampling bias and thus reduced overfitting, which in turn improves the transferability of the models in space and time (Boria et al., 2014). A limiting factor of distribution modelling of introduced seaweeds for future climate predictions is the availability of future climate predictions of other abiotic factors, like pH and phosphor, that have been shown to be important predictors of seaweed distributions (Verbruggen et al., 2013).

The turnover maps show that the species composition in certain places will be altered, even when there is a limited increase in the total number of introduced species. If one or a few 'leverage species' become suitable or unsuitable this may result in sweeping community-level changes (Harley et al., 2006).

Conclusion

Distribution modelling of invasive seaweeds is a challenging task. In this study we showed that using coastal background, spatial thinning and an ensemble of models with quadratic features results in transferable distribution models. However, predicting the invasion through time can yield poorly performing models when the known distribution records don't reflect the environmental niche of the species. Change and turnover maps combined with an assessment of the uncertainty therein are valuable tools. They allow for a cost-effective monitoring of coastlines, as not all European coastlines will be evenly impacted by climate change.

Acknowledgements

The research was carried with financial support from the ERANET INVASIVES project (EU FP7 SEAS-ERA/INVASIVES SD/ER/010) and financial support provided by VLIZ as part of the Flemish contribution to the LifeWatch ESFRI. The authors would like to thank the INVASIVES consortium members for contributing data and providing suggestions; Sara Martinez and Zoë De Corte for collecting data and exploratory work.

Supporting information

In this supporting information we first aim to build transferable SDMs by selecting the best modelling choices. In order to be able to select transferable models the cross-validation (CV) splitting approaches have to be compared. Of the three cross-validation data splitting approaches (CV) the random splitting approach yielded the highest values across the evaluation metrics used: AUC, kappa and H-measure (Table S1). On the other hand, the temporal splitting approach resulted in the lowest evaluation metrics for all species, with some of the models becoming indistinguishable from random models. The random CV resulted in occurrence training and test sets with very similar densities in environmental space, that are dissimilar from the background test points (Fig. S1). The spatial CV has occurrence training and test records that are both dissimilar from each other and from the background points. Lastly, with the temporal splitting approach both the occurrence training and test sets and the background test points are very similar. From this point onward options were only evaluated using the spatial CV.

None of the two thinning distances explored (30 versus 100 km) performed better for all species (Table S2). Therefore the thinning procedure was kept species-specific. The thinning distances used were 30 km for *G. turuturu* and *S. muticum* and 100 km for the other species.

Table S3 shows that models built with coastal background have higher AUC values. They are the most transferable for all features, species and algorithms. As the Surface Range Envelope algorithm only uses occurrence data to build the model, the resulting AUC values are the same for both types of background.

Including quadratic features results in a higher transferability of the models for both MaxEnt and GLM (Table S3). Regarding the different algorithms tested, although generalized linear models consistently have a high AUC (Table S3A), other coastal background algorithms sometimes perform better than GLM depending on the species. In addition, MaxEnt coastal models with quadratic features perform somewhat similarly to coastal GLMs with quadratic features. The ensemble model built using all the algorithms, with the coastal background and quadratic features (for MaxEnt and GLM), generally performs well. The results for the other evaluations metrics show the same trends as those for AUC described here (Table S3B and C).

Table S1. Model performance for different algorithms and cross-validation (CV) data splitting approaches for all species. Values in red indicate high values while values in white indicate low values. A coastal background was used for all models and MaxEnt and GLM algorithms were performed with quadratic features. Thinning distances used for the different species were 100, 100, 30, 30 and 100 km, respectively. The three evaluation metrics used are AUC (A), kappa (B) and the H-measure (C).

A. AUC

CV	Algorithm	*D. cyanoloma*	*C. fragile*	*G. turuturu*	*S. muticum*	*U. pinnatifida*
Spatial	GLM	0.89	0.769	0.86	0.866	0.899
	MaxEnt	0.764	0.765	0.682	0.862	0.755
	RF	0.713	0.63	0.641	0.652	0.764
	SRE	0.727	0.796	0.743	0.613	0.813
Year	GLM	0.576	0.598	0.798	0.593	0.607
	MaxEnt	0.502	0.593	0.663	0.578	0.556
	RF	0.46	0.568	0.749	0.558	0.529
	SRE	0.567	0.588	0.743	0.621	0.626
Random	GLM	0.964	0.908	0.938	0.939	0.948
	MaxEnt	0.949	0.91	0.937	0.94	0.935
	RF	0.915	0.908	0.949	0.957	0.931
	SRE	0.893	0.835	0.805	0.848	0.86

B. Kappa

CV	Algorithm	*D. cyanoloma*	*C. fragile*	*G. turuturu*	*S. muticum*	*U. pinnatifida*
Spatial	GLM	0.636	0.345	0.657	0.667	0.702
	MaxEnt	0.424	0.345	0.357	0.638	0.471
	RF	0.182	-0.024	-0.086	0.039	-0.038
	SER	0.455	0.595	0.486	0.226	0.615
Year	GLM	0.067	0.088	0.378	0.151	0.058
	MaxEnt	0.067	0.076	0.324	0.196	0.084
	RF	-0.067	0.069	0.135	0.063	0.027
	SER	0.133	0.176	0.486	0.242	0.254
Random	GLM	0.857	0.675	0.78	0.725	0.824
	MaxEnt	0.786	0.675	0.805	0.725	0.765
	RF	0	0.147	0.146	0.469	0.059
	SER	0.786	0.669	0.61	0.689	0.721

C. H-measure

CV	Algorithm	*D. cyanoloma*	*C. fragile*	*G. turuturu*	*S. muticum*	*U. pinnatifida*
Europe	GLM	0.539	0.301	0.492	0.509	0.608
	MaxEnt	0.354	0.299	0.185	0.506	0.296
	RF	0.255	0.1	0.132	0.094	0.28
	SER	0.391	0.394	0.381	0.093	0.454
Year	GLM	0.155	0.084	0.39	0.097	0.108
	MaxEnt	0.128	0.079	0.232	0.105	0.109
	RF	0.12	0.036	0.288	0.032	0.04
	SER	0.024	0.056	0.281	0.092	0.111
Random	GLM	0.853	0.542	0.756	0.648	0.752
	MaxEnt	0.884	0.542	0.732	0.648	0.705
	RF	0.783	0.529	0.736	0.735	0.694
	SER	0.768	0.449	0.544	0.531	0.61

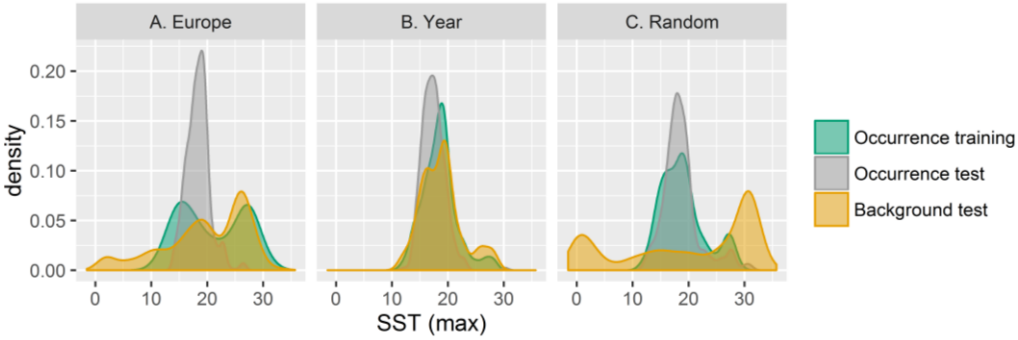


Figure S1. Distribution of occurrences (training and test) and background test for the maximum sea surface temperature for *Sargassum muticum* for the three different splitting approaches: Europe (A), year (B) and random (C).

Table S2. Model performance for the two thinning distances, 30km and 100km for the three different metrics: AUC (A), kappa (B) and H-measure (C). GLM and MaxEnt models were built with quadratic features. The best thinning distance is marked in green.

A. AUC

Algorithm	*D. cyanoloma*		*C. fragile*		*G. turuturu*		*S. muticum*		*U. pinnatifida*	
	30 km	100 km	30 km	100 km	30 km	100 km	30 km	100 km	30 km	100 km
GLM	0.881	0.895	0.766	0.769	0.86	0.665	0.866	0.688	0.907	0.899
MaxEnt	0.728	0.764	0.82	0.765	0.682	0.631	0.862	0.683	0.779	0.755
RF	0.705	0.713	0.725	0.63	0.641	0.59	0.652	0.566	0.817	0.764
SRE	0.727	0.727	0.805	0.796	0.743	0.75	0.613	0.593	0.827	0.813

B. Kappa

Algorithm	*D. cyanoloma*		*C. fragile*		*G. turuturu*		*S. muticum*		*U. pinnatifida*	
	30 km	100 km	30 km	100 km	30 km	100 km	30 km	100 km	30 km	100 km
GLM	0.636	0.636	0.278	0.345	0.657	0.443	0.667	0.211	0.692	0.702
MaxEnt	0.394	0.424	0.487	0.345	0.357	0.314	0.638	0.166	0.51	0.471
RF	0.212	0.182	0.216	-0.024	-0.086	-0.014	0.039	-0.005	0.038	-0.038
SRE	0.455	0.455	0.612	0.595	0.486	0.5	0.226	0.184	0.663	0.615

C. H-measure

Algorithm	*D. cyanoloma*		*C. fragile*		*G. turuturu*		*S. muticum*		*U. pinnatifida*	
	30 km	100 km	30 km	100 km	30 km	100 km	30 km	100 km	30 km	100 km
GLM	0.534	0.539	0.261	0.301	0.492	0.206	0.509	0.169	0.599	0.608
MaxEnt	0.312	0.354	0.395	0.299	0.185	0.158	0.506	0.168	0.324	0.296
RF	0.25	0.255	0.195	0.1	0.132	0.083	0.094	0.036	0.383	0.28
SRE	0.389	0.391	0.418	0.394	0.381	0.399	0.093	0.076	0.512	0.454

Table S3 Overview of the effect of the different modelling choices on the model performance. Columns are divided by species and background type (coastal and global background), rows represent modelling algorithms and feature types (linear or quadratic), and the values are the performance metrics area under the curve (A), kappa (B) and H-measure (C). A higher value (red) indicates a high transferability, while low values (white) indicate a poor performance. Thinning distances used for the different species were 100, 100, 30, 30 and 100 km, respectively. Ensemble models for the different species were built with the options selected with an asterisk (*).

A. AUC

	D. cyanoloma		*C. fragile*		*G. turuturu*		*S. muticum*		*U. pinnatifida*	
Algorithm	Coastal	Global	Coastal	Global	Coastal	Global	Coastal	Global	Coastal	Global
GLM Q	0.895	0.768	0.769	0.666	0.86	0.383	0.866	0.771	0.899	0.827
GLM L	0.651	0.647	0.49	0.351	0.529	0.329	0.505	0.492	0.56	0.508
MaxEnt Q	0.764	0.662	0.765	0.536	0.682	0.379	0.862	0.762	0.755	0.529
MaxEnt L	0.661	0.632	0.485	0.359	0.527	0.386	0.51	0.484	0.564	0.487
RF	0.713	0.655	0.63	0.498	0.641	0.46	0.652	0.392	0.764	0.745
SRE	0.727	0.727	0.796	0.797	0.743	0.743	0.613	0.614	0.813	0.808
Ensemble	0.894		0.831		0.802		0.843		0.885	

B. Kappa

	D. cyanoloma		*C. fragile*		*G. turuturu*		*S. muticum*		*U. pinnatifida*	
Algorithm	Coastal	Global	Coastal	Global	Coastal	Global	Coastal	Global	Coastal	Global
GLM Q	0.636	0.515	0.345	0.297	0.657	-0.071	0.667	0.286	0.702	0.538
GLM L	0.455	0.485	0.259	-0.012	0.157	-0.229	0.099	0.077	0.327	0.115
MaxEnt Q	0.424	0.394	0.345	0.136	0.357	-0.043	0.638	0.358	0.471	0.135
MaxEnt L	0.455	0.394	0.253	0.003	0.171	-0.057	0.102	0.047	0.317	0.106
RF	0.182	0.061	-0.024	0	-0.086	-0.143	0.039	-0.031	-0.038	0.317
SRE	0.455	0.455	0.595	0.59	0.486	0.486	0.226	0.226	0.615	0.625
Ensemble	0.545		0.463		0.4		0.461		0.519	

C. H-measure

	D. cyanoloma		*C. fragile*		*G. turuturu*		*S. muticum*		*U. pinnatifida*	
Algorithm	Coastal	Global	Coastal	Global	Coastal	Global	Coastal	Global	Coastal	Global
GLM Q	0.539	0.367	0.301	0.151	0.492	0.068	0.509	0.34	0.608	0.366
GLM L	0.256	0.239	0.065	0.133	0.091	0.083	0.033	0.074	0.119	0.043
MaxEnt Q	0.354	0.251	0.299	0.056	0.185	0.057	0.506	0.339	0.296	0.052
MaxEnt L	0.293	0.25	0.068	0.133	0.096	0.065	0.034	0.074	0.123	0.039
RF	0.255	0.17	0.1	0.01	0.132	0.027	0.094	0.067	0.28	0.257
SRE	0.391	0.391	0.394	0.394	0.381	0.381	0.093	0.093	0.454	0.454
Ensemble	0.562		0.451		0.414		0.478		0.526	

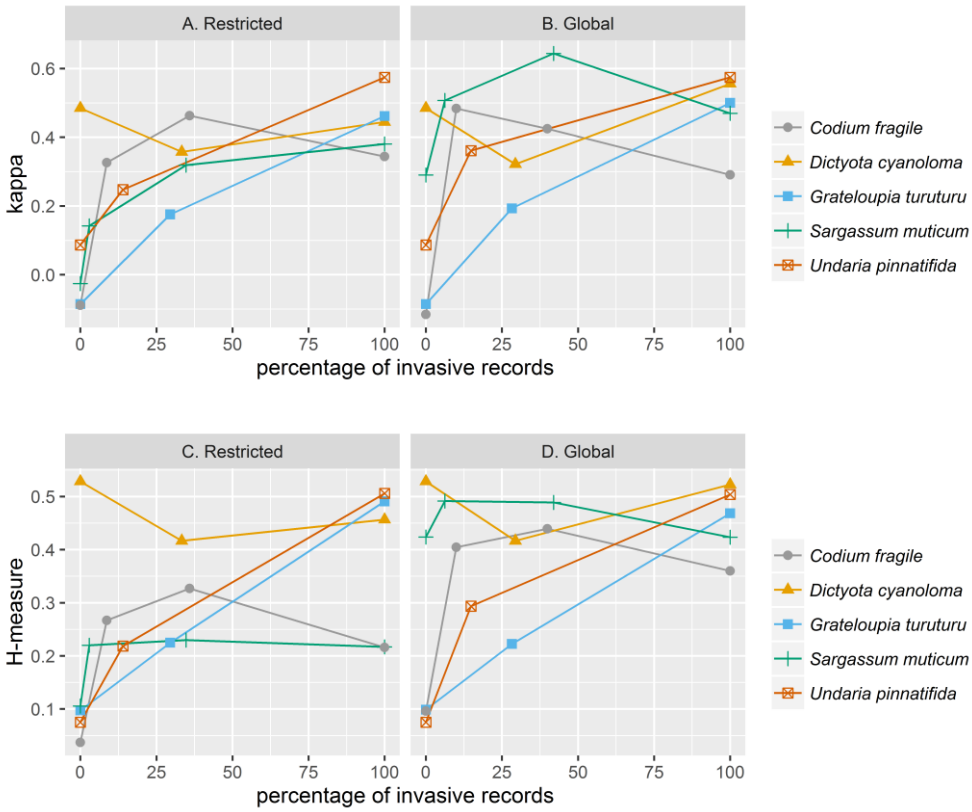


Figure S2. Evolution of kappa (A, B) and H-measure (C, D) at different points along the invasive history. The left figure (A, C) shows the values for the restricted scenario, in which at T1 only native records are used for model fitting. Subsequent time points use both native and European records. For the right figure (B, D) all occurrence records known at the specific years are used for modelling. The x-axis represents the percentage of invasive records included to build the model with the total number of records included in the last model as 100%.

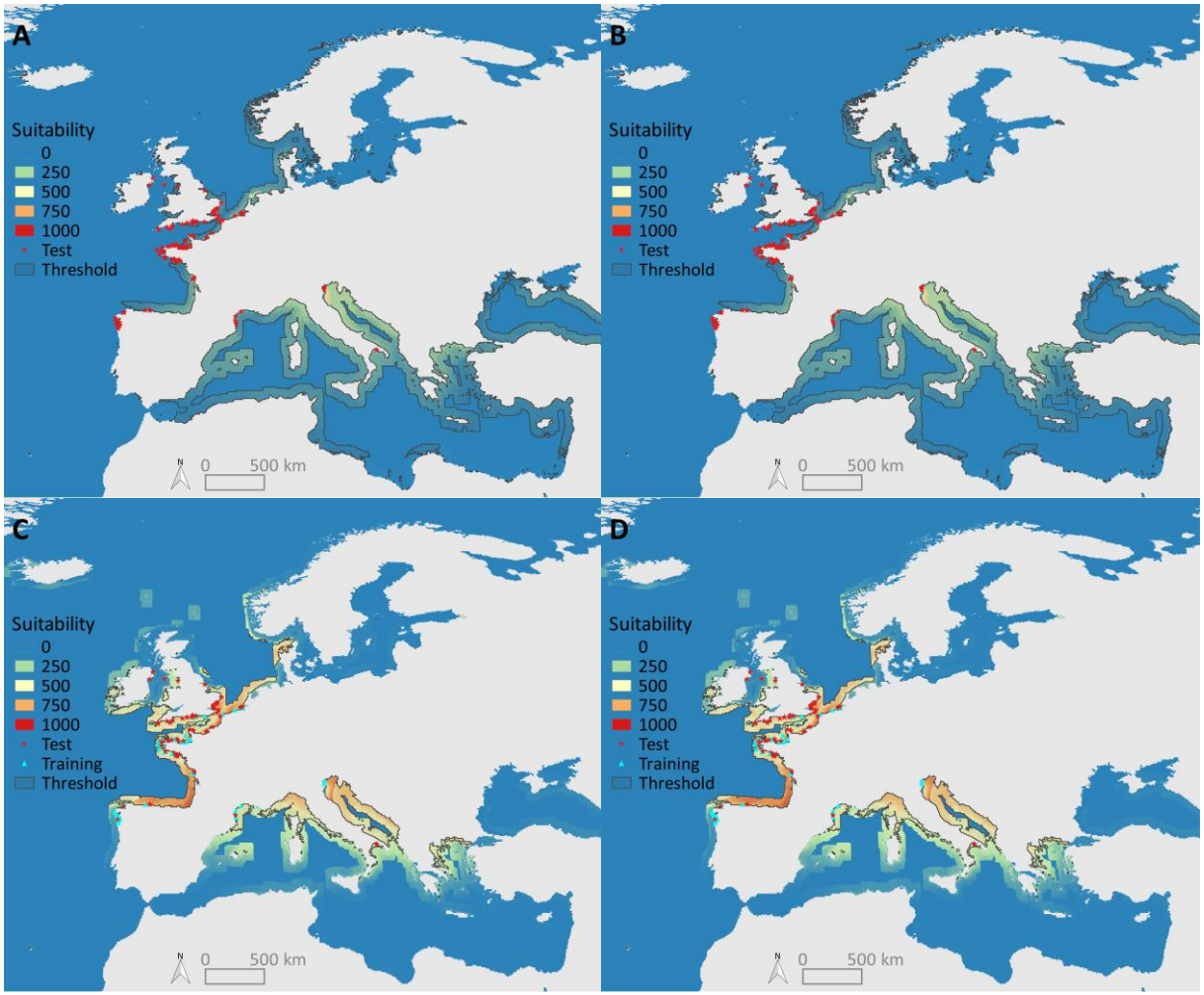


Figure S3. European predictions of suitable areas for *Grateloupia turuturu*. Red stars are locations used as test occurrences and the training records are in cyan triangles. The records for fitting the models are: A) only native records, B) native records and all invasive records known before the introduction of *G. turuturu* in Europe, C) native records and all European records known in the year 2000 (T3) and D) native records and all invasive records known in 2000. A) and C) represent models from the restricted scenario while B) and D) are models from the global scenario. For further information about the number of records included in each model we refer to Table 2.

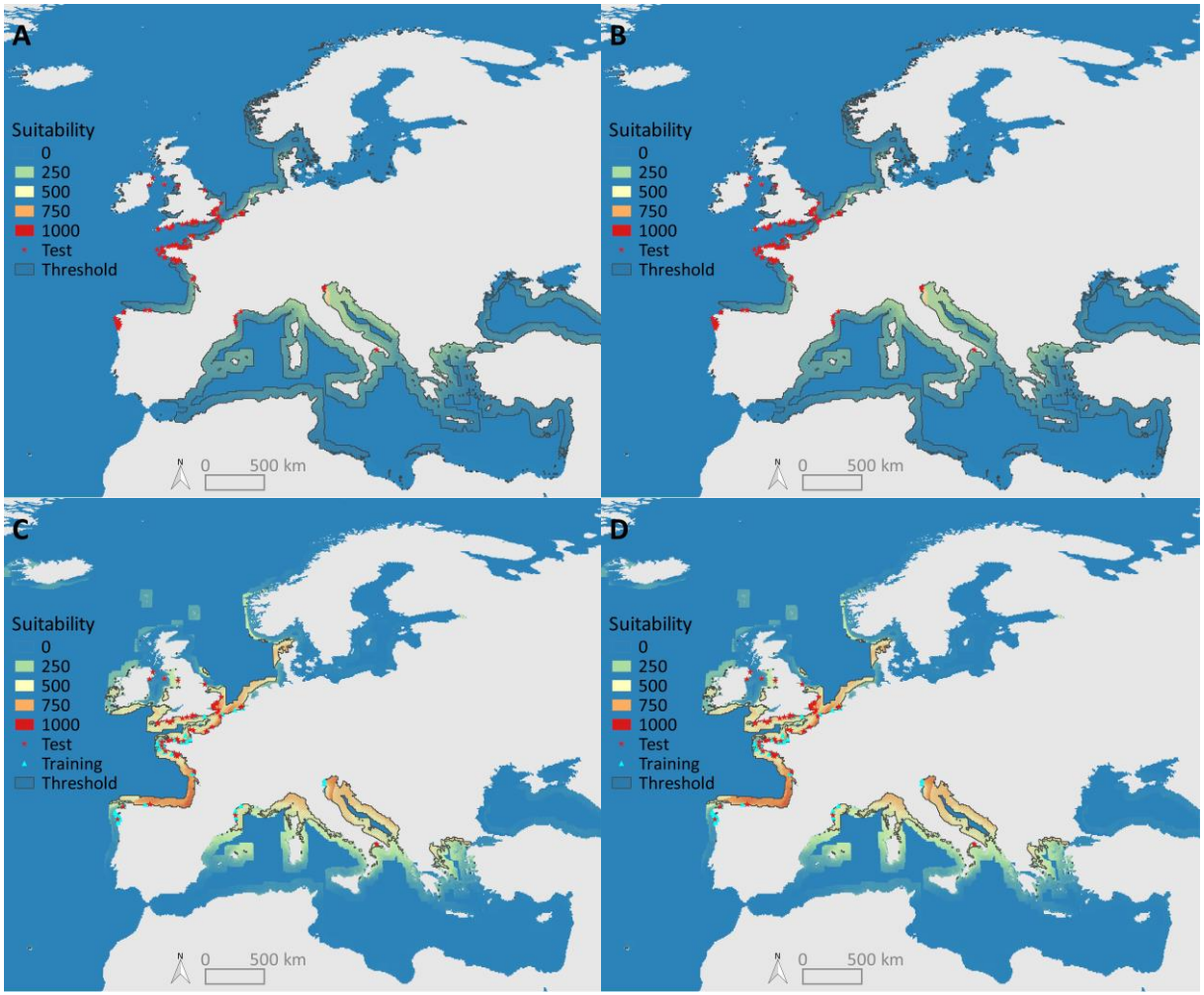


Figure S4. European predictions of suitable areas for *Undaria pinnatifida*. Red stars are locations used as test occurrences and the training records are in cyan triangles. The records for fitting the models are: A) only native records, B) native records and all invasive records known before the introduction of *U. pinnatifida* in Europe, C) native records and all European records known in the year 2000 (T3) and D) native records and all invasive records known in 2000. A) and C) represent models from the restricted scenario while B) and D) are models from the global scenario. For further information about the number of records included in each model we refer to Table 2.

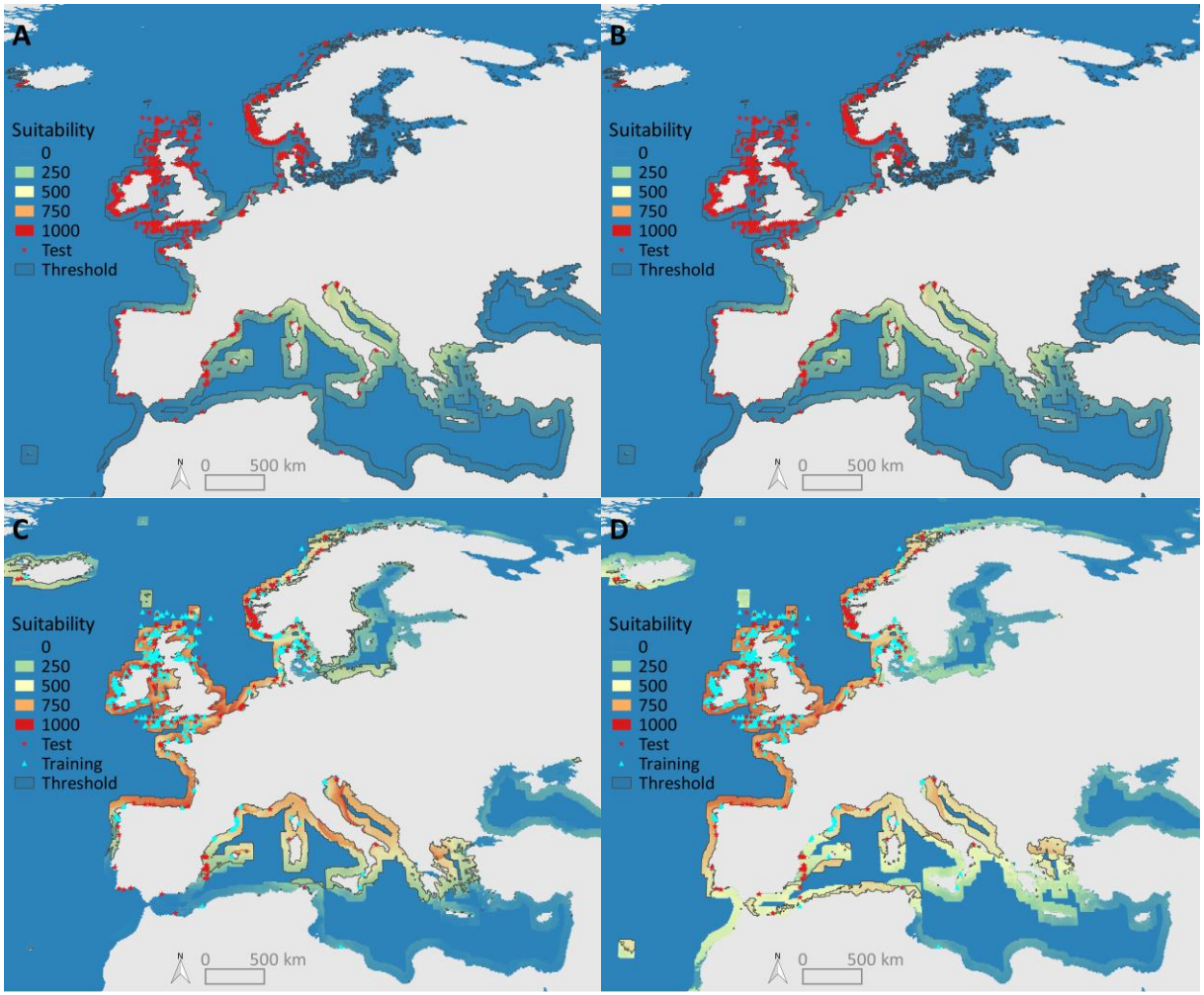


Figure S5. European predictions of suitable areas for *Codium fragile* subsp. *fragile*. Red stars are locations used as test occurrences and the training records are in cyan triangles. The records for fitting the models are: A) only native records, B) native records and all invasive records known before the introduction of *C. fragile* in Europe, C) native records and all European records known in the year 1990 (T4) and D) native records and all invasive records known in 1990. A) and C) represent models from the restricted scenario while B) and D) are models from the global scenario. For further information about the number of records included in each model we refer to Table 2.

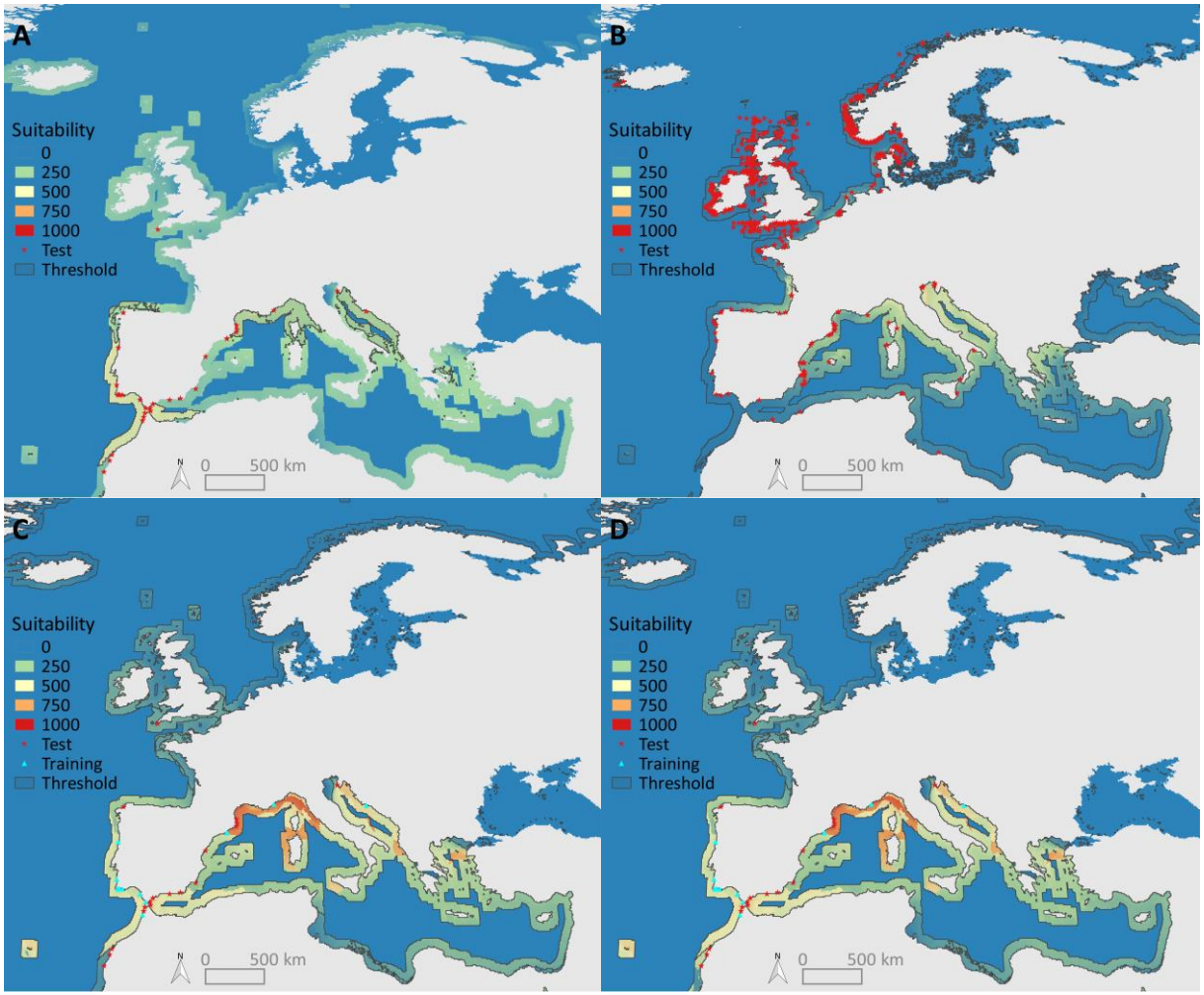


Figure S6. European predictions of suitable areas for *Dictyota cyanoloma*. Red stars are locations used as test occurrences and the training records are in cyan triangles. The records for fitting the models are: A) only native records, B) native records and all invasive records known before the introduction of *D. cyanoloma* in Europe, C) native records and all European records known in the year 2010 (T3) and D) native records and all invasive records known in 2010. A) and C) represent models from the restricted scenario while B) and D) are models from the global scenario. For further information about the number of records included in each model we refer to Table 2.

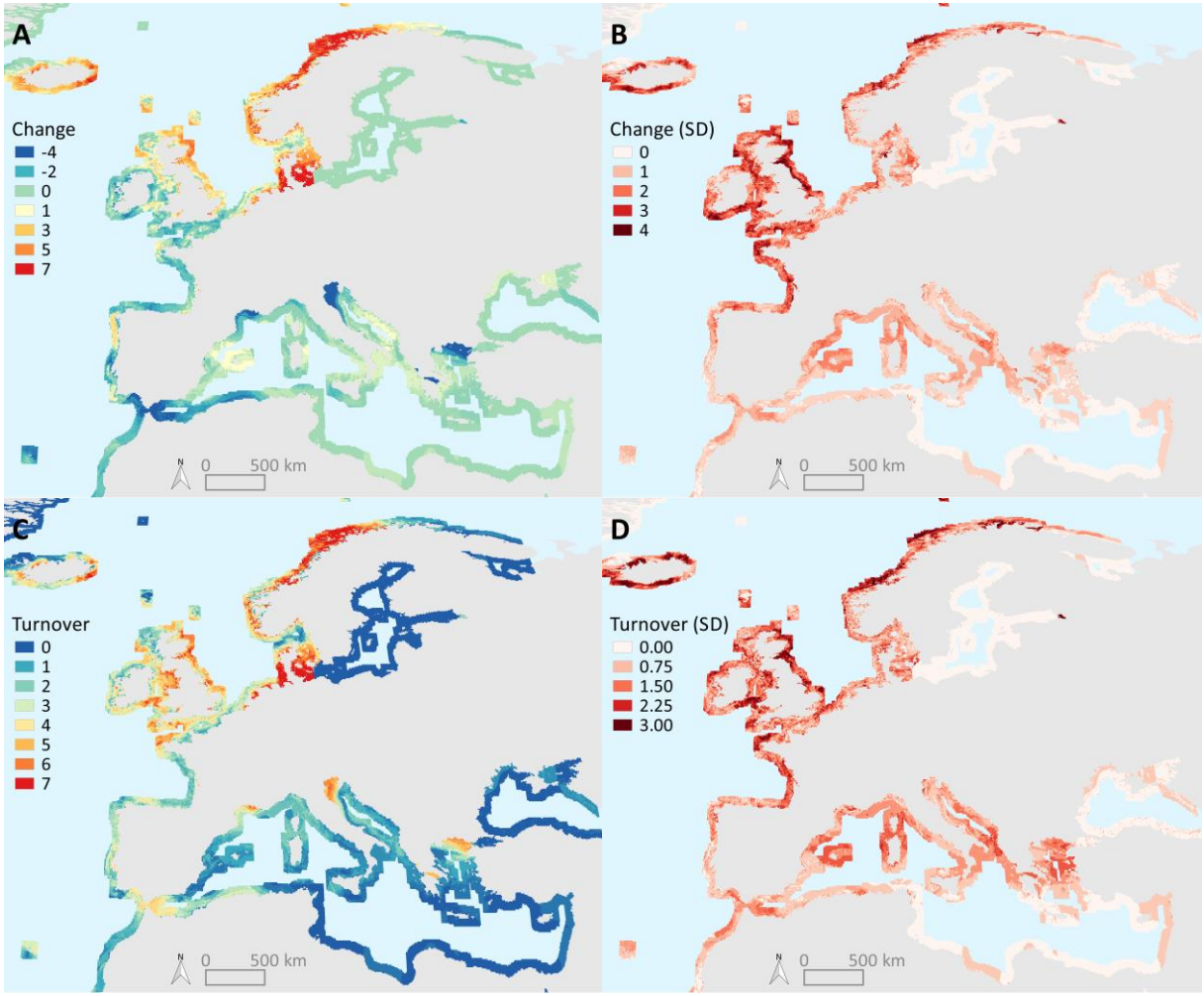


Figure S7. The change in number and turnover of introduced seaweeds predicted by 2100 under the IPCC climate change scenario A1B. The top left map (A) shows the mean difference in number of introduced species between the current climate and climate change scenario A1B from SDMs build for 15 species. The top right map (B) indicates the standard deviation in model predictions when using the mean, minimum or maximum sea surface temperature as one of the four predictors for building the distribution models for each species. The bottom left (C) and right (D) maps indicate the mean and standard deviation of the turnover of introduced seaweeds.

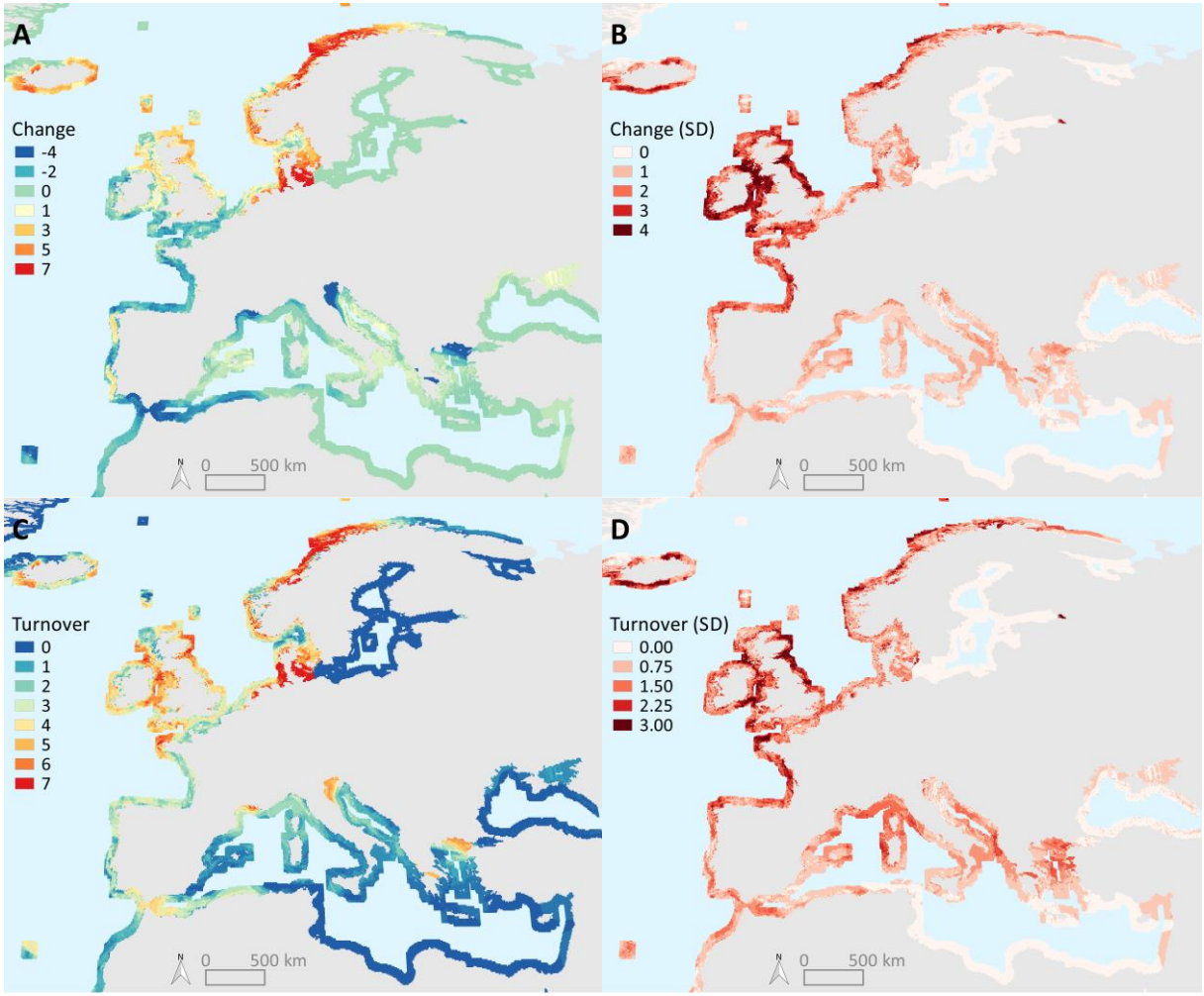


Figure S8. The change in number and turnover of introduced seaweeds predicted by 2100 under the IPCC climate change scenario A2. The top left map (A) shows the mean difference in number of introduced species between the current climate and climate change scenario A2 from SDMs build for 15 species. The top right map (B) indicates the standard deviation in model predictions when using the mean, minimum or maximum sea surface temperature as one of the four predictors for building the distribution models for each species. The bottom left (C) and right (D) maps indicate the mean and standard deviation of the turnover of introduced seaweeds.

Chapter 8

General discussion

The question of how plants and animals are distributed on Earth in space and time has fascinated biologists for a long time. The structure in the patterns observed inspired biogeographers and ecologists to seek explanations (Guisan & Thuiller, 2005). This fascination has acquired an additional significance in the 21th century now that it becomes evident that the global climate is changing at an unprecedented scale. Climate exerts a dominant control over the natural distribution of species (Pearson & Dawson, 2003) and changes in climatic conditions are currently altering ranges of terrestrial and marine species (Walther et al., 2002; Parmesan & Yohe, 2003; Perry, 2005; Chen et al., 2011). Without intending to downplay the effect of global climate change on terrestrial environments there is strong evidence that especially coastal marine ecosystems are affected by global climate change (Harley et al., 2006; Doney et al., 2012). It has been repeatedly suggested that the relentless anthropogenic pressure by pollution, eutrophication, overexploitation, habitat destruction and human-mediated exchanges of species has important cumulative effects on marine coastal environments (Strain et al., 2014). These anthropogenic stressors have the capacity to undermine the natural resilience of coastal ecosystems and exacerbate the effect of global climate change. For example, the distribution ranges of many intertidal species have shifted by as much as 50 km per decade, which is much faster than most recorded shifts of terrestrial species (Helmuth et al., 2006). Understanding the response of coastal marine ecosystems and predicting the potential impacts of a changing environment as a consequence of global change is therefore important and timely.

Species distribution models that correlate the distributions of a species with climate variables or through an understanding of species' physiological responses to climate change can present a valuable tool to understand the environmental conditions governing a species distribution range and to predict changes in the range distribution as a consequence of climate change (Pearson & Dawson, 2003). Despite this potential, species distribution modelling in the marine environment is less studied than for terrestrial ecosystems (Robinson et al., 2011).

In this thesis we tried to push marine species distribution modelling by making distribution and environmental data more accessible to the end-user (Chapter 2, 3) and to advance our understanding of predictor relevance (Chapter 4). To achieve the latter we compiled a benchmark dataset which should facilitate future comparative studies of various aspects of the marine species distribution modelling process. In addition we applied these findings to test cases consisting of invasive seaweeds along European coasts.

Below I discuss the main findings in a broader perspective and provide perspectives on how to improve species distribution modelling in the future.

Data

Data about marine species and their distributions is an essential part of contemporary biological studies. The use of species-level data is required to get an understanding of the spatial pattern of marine ecosystems, their evolution, and how they respond to environmental change (Grassle, 2000).

An ever increasing amount of taxonomic and distributional information is made available through web databases (see Table 1 for a small overview) like World Register of Marine Species (worms.org), Ocean Biogeographic Information System (iobis.org), Global Biodiversity Information Facility (gbif.org), Encyclopedia Of Life (eol.org), etc. Next to these general databases a wide variety of specialized websites exist for e.g. introduced species (DASIE, www.europe-aliens.org), fish (FishBase, fishbase.org), coral reefs (ReefBase, reefbase.org), algae (Algaebase, algaebase.org), etc. Moreover, multiple museum and herbarium collections have been released online e.g. Macroalgal Herbarium Portal (macroalgae.org), Australia's Virtual Herbarium (avh.chah.org.au), Natural History Museum London (nhm.ac.uk) and Muséum National d'Histoire Naturelle (mnhn.fr). Finally, citizen science initiatives contribute significantly towards international biodiversity monitoring (Chandler et al., 2016). While most initiatives are geared towards terrestrial species, some marine citizen projects are well established (Edgar & Stuart-Smith, 2009; Theobald et al., 2015; Edgar et al., 2016). Note that several of these databases exchange data, for example data from Algaebase and FishBase is shared with WoRMS and the Natural History Museum London collection data is uploaded to GBIF. Bingham et al. (2017) give a broad overview of this global landscape of biodiversity databases, projects and datasets, and the relations between these elements.

Data from these databases have been used for a wide variety of applications. For instance, records from OBIS have in the last two years been used to model shifts in distributions of fish species (Fogarty et al., 2017), classify degrees of species commonness (Coro et al., 2015), identify species richness patterns (Chaudhary et al., 2016; Ma et al., 2017), describe the biodiversity and biogeography of intertidal communities (Griffiths & Waller, 2016), identify biases in biodiversity (Higgs & Attrill, 2015) and study marine spatial planning (Caldow et al., 2015; Geijzendorffer et al., 2016). In Box 1 we give some modelling perspectives for new data in these public databases.

Table 1. Overview of the online databases with taxonomic and biogeographic data used in this thesis. Dark grey indicates the focus of the database as a source for taxonomy, biogeography, traits and herbarium data. Some of the database only focus on the marine environment (dark grey) while others collect data from both the marine and terrestrial environment. While citizen science data ends up in multiple of the listed databases only Reef Life Survey is focused on this. Although some databases have geographic focus, data from sampling campaigns in other parts of the world often ends up in these databases.

Online database	Taxonomy	Biogeography	Traits	Herbarium	Only marine	Citizen science	Taxonomic focus	Geographic focus
Algaebase (algaebase.org)	■						Algae	World
Australia's Virtual Herbarium (avh.chah.org.au)				■			Plants, algae and fungi	Australia
Encyclopedia Of Life (eol.org)	■		■					World
FishBase (fishbase.org)	■	■	■				Fish	World
Global Biodiversity Information Facility (gbif.org)		■						World
Macroalgal Herbarium Portal (macroalgae.org)				■	■		Algae	World
Muséum National d'Histoire Naturelle (mnhn.fr)		■		■				France
Natural History Museum London (nhm.ac.uk)		■		■				United Kingdom
Ocean Biogeographic Information System (iobis.org)		■			■			World
Reef Life Survey (reeflifesurvey.com)		■			■	■	Reefs	World
ReefBase (reefbase.org)		■			■			World
World Register of Marine Species (worms.org)	■		■		■			World

All of the above mentioned data sources were used in one or more of the studies performed for this thesis. However, we encountered, several limitations related to the available information on the taxonomy and distribution of species. Caveats became most apparent while characterizing trends in introduced seaweeds in Chapter 5. Especially taxonomic problems were rife as for a large percentage of the species the taxonomic identity could not be accurately assessed. Furthermore, the introduction itself was uncertain or the native area of the species was unknown. Taxonomic issues and uncertainty with respect to the introduction history can often be resolved by performing genetic analyses of in situ and herbarium samples (McIvor et al., 2001; Verbruggen et al., 2007; Provan et al., 2008; Steen et al., 2017).

Box 1. New data for predictive modelling

As shown in previous chapters, taxonomic databases such as the World Register of Marine Species (WoRMS) and Encyclopedia Of Life (EOL) and biogeographic databases like the Ocean Biogeographic Information System (OBIS) and the Global Biodiversity Information Facility (GBIF) are of great importance for marine predictive modelling. The inclusion of new types of data will allow for new applications and improvements of the current applications.

Regarding taxonomic databases, specific attention has been given towards the inclusion of species' traits information. The development of a marine species traits vocabulary was one of the tasks of the European Marine Observation Data Network Biology project (www.emodnet-biology.eu). As proposed in Chapter 2, traits from WoRMS can be very valuable for creating new quality control procedures for biogeographic databases, by allowing for example flagging a coastal species observed in the open ocean. When traits can be linked to a species' distribution, this information can be used to study the relationship between traits and predictors, background data, algorithms and model complexity. In Chapter 7 we identified one such relationship between seaweeds and background data concluding that the usage of coastal background data results in a higher transferability of the models. Another unreported aspect concerns the relationship between different species, e.g. community membership, food web structure and cryptic species complexes. Having distributional information on other species that are part of the same community as the species being modelled allows for community-level approaches to species distribution modelling, which can, especially for rare species, result in improved model predictions (Elith et al., 2006; Hui et al., 2013; Madon et al., 2013; Harris, 2015).

For the biogeographic databases, OBIS has proposed to store in its database the sampling methodology, animal tracking and telemetry data, biological measurements (e.g., body length, percent live cover, ...) and environmental measurements such as nutrient concentrations, sediment characteristics or other abiotic parameters measured during sampling (De Pooter et al., 2017). While these data run risk to be highly heterogeneous, they potentially allow for studies of the link between the presence and abundance of species and the abiotic factors available for the species at the moment of the sampling event. The additional information on sampling methodology could be valuable metadata for the assessment of the fitness for use of the sample for a specific use case.

However, a DNA reference framework is still missing for most macroalgae and for some species their origin will remain unknown. We also find a geographic signature with respect to taxonomic uncertainty, which is more common in Macaronesia and the Mediterranean Sea compared to the NE Atlantic. This difference is probably not due to taxonomic effort which we would expect to be more or less equal across these regions. In our opinion the higher percentage of species with tropical affinities

contributes significantly to taxonomic uncertainty. Tropical seaweeds in particular are poorly characterized with many 'species' being reported from all major ocean basins. In many cases these so-called pantropical species turn out to represent complexes of cryptic species. The individual species have predominantly, but not always, relatively small and well-defined ranges (e.g. Payo et al. 2013, Silberfeld et al. 2013, Vieira et al. 2017). Reports of widespread tropical species in the Mediterranean Sea or Macaronesia are therefore very difficult to interpret correctly. The species could after all be native and if introduced their area of origin is difficult to assess.

While developing the MarineSPEED benchmark dataset (Chapter 4) we specifically aimed to avoid species with the above mentioned taxonomical issues by selecting well-studied species. During this process we identified the need for an indicator of taxonomic reliability on two levels. On a species level it would be useful to indicate whether a species is part of a group of cryptic species or whether it potentially is a synonym of another species. Additionally, an indication of the taxonomic reliability on specimen-level based on the identification source could prove valuable as we expect that for many species identification based on genetic information is more reliable than identification based on visual information.

In Chapter 2 we aimed to alleviate the problems associated with publicly available databases containing distribution records from various sources. While OBIS records and their quality control flags have been made available through the *robis* R package (Provoost et al., 2016), the records available from OBIS and other species distribution databases, display significant differences in the quality and abundance of information on different parts of the world. By collecting data for a benchmark dataset and for modelling introduced seaweeds we identified multiple issues. The first issue consists of the uneven distribution of researchers throughout the marine world leading to clear biases, linked to the distribution of economic wealth. Secondly, even similar research effort can lead to pronounced differences in the abundance of distribution records. For instance, at a European scale we noted marked differences in availability of distribution records of macroalgae. Distribution records from France and the Mediterranean Sea are significantly less represented than records from the United Kingdom. We additionally encountered a paucity of distribution records in countries from the North Western Pacific (e.g. Japan, Russia, China and South-Korea), the native area of many introduced seaweed species. Australia on the other side is much better represented. For our analyses of trends of introduced seaweeds in Europe (Chapter 5), the 4,900 distribution records used were all sourced from literature and from participants of the INVASIVES project.

These records are by and large not available from public databases, indicating that there is a lot more data available. Gathering new distribution records and georeferencing from literature records is especially tedious as sampling locations are often only mentioned in the text or on a map and not as coordinates. The lack of distribution records refrained us from applying species distribution modelling techniques for the species identified in aquaria (Chapter 6). For future publications, this problem can be avoided by adopting a paper submission policy which requires the submission of distribution information to OBIS or GBIF just like is currently required for DNA sequences to GenBank (Costello et al., 2013).

The promotion of current and new citizen science initiatives can be invaluable if we want to improve the number and quality of the available distribution information. New citizen science initiatives for extracting distribution information from literature sources combined with text mining could allow for the inclusion of literature records in public databases. However special care should be taken to avoid errors. A common error includes, dating distribution records as the publication date, which is generally not the same as the date the records were sampled. Moreover, one should avoid including fossil specimens (e.g. Foraminifera) in databases of contemporary distributions. Although citizen science projects are very appealing because of the relatively low costs involved and the resulting data has similar accuracy as data from other sampling campaigns (Edgar & Stuart-Smith, 2009), it can only be used as a supplement to funded research as the resulting data, especially from observations, is biased towards larger species.

In Chapter 2 we reported the introduction of quality control flags for (Eur)OBIS. Such quality flags are a useful tool for feedback towards data publishers and allow for filtering of records according to the needs of the study envisioned by the researchers. This process could be improved by adding additional flags to datasets based on the study design and by grouping quality control flags into useful groups based on specific criteria, e.g. abundance-related or time-related criteria. A trade-off, however, exists between the stringency of the quality control and obtaining sufficient data for modelling. While some progress has been made in fuzzy and Bayesian modelling techniques, a further development of these incorporating the uncertainty in distribution records could allow for the inclusion of the available information while taking into account the quality of the data (Mouton et al., 2009; Hattab et al., 2013; Costa et al., 2015; Hamilton et al., 2015; Golding & Purse, 2016).

The currently available datasets with global environmental data such as WorldClim (Hijmans et al., 2005) and Bio-ORACLE (Tyberghein et al., 2012) have proven their

value in numerous species distribution modelling studies. Nevertheless, it should be acknowledged that these predictors often only provide a rough approximation of the abiotic factors controlling a species' distribution and the usage of other predictors may lead to better models. Recently, Title and Bemmels (2017) identified an additional set of 18 new terrestrial climatic and topographic variables that are likely to have direct relevance to ecological or physiological processes determining terrestrial species distributions. Similarly, a new extended version of Bio-ORACLE is being created which addresses the need for high resolution benthic layers for modelling marine benthic species (Davies & Guinotte, 2011; Reiss et al., 2015; Boavida et al., 2016), which make up the bulk of the marine diversity. For example, the exploration of deep cryptic refugia for marine benthic species, is suboptimal when distributions are modelled using surface data only (Graham et al., 2007b; Assis et al., 2016). Besides benthic data for temperature, silicate, salinity, primary production, carbon phytoplankton biomass, nitrate, light, phosphate, iron, dissolved oxygen, current velocity and chlorophyll, data on sea surface temperature, sea ice and bathymetry is provided. All layers will be generated for the current and future climate and be made available through the *sdmpredictors* R package (Chapter 3).

While these new benthic layers in Bio-ORACLE2 will no doubt open up new possibilities for improvements in marine benthic species distribution modelling it will render the approach used in Chapter 4, to model the relevance of predictors by fitting distribution models for all combinations of predictors, computationally unfeasible as the number of combinations of predictors increases non-linearly. Using the variable importance from a limited number of predictor sets (Brandt et al., 2017), combined with an analysis of synergistic and antagonistic interaction effects, may allow for a considerable reduction of this computational burden.

Predictive models and uncertainty

In the previous chapters different models were created related to the distribution of marine species. In Chapter 2 we used a simple but effective outlier detection model based on the median absolute distance and the interquartile range to detect environmental and geographical outliers in distribution records published on OBIS. The relevance of available marine predictors was modelled by ranking the performance of predictors in species distribution models for various combinations using the MarineSPEED benchmark dataset in Chapter 4. Trends in the number of introduced seaweeds in Europe were modelled using logistic regression in Chapter 5. In Chapter 6 we estimated total species richness and created thermal niche models in order to assess the risk for seaweed introductions in Europe due to aquarium

trade. Lastly, in Chapter 7 we modelled the distribution and niche of invasive and introduced seaweeds in Europe.

While all of the above approaches provide models fit for the purpose, the need for the assessment and visualization of the uncertainty is apparent. In Chapter 4 variation in predictor relevance, and thus uncertainty, was visualized by using box plots of the results from various combinations of modelling setups. The uncertainty in the midpoint of the best fitting logistic curve for introduced seaweeds was assessed using bootstrapping (Chapter 5, Efron and Tibshirani 1986). Considerable uncertainty is involved when estimating total species richness in aquarium trade and on live rock which is visualized by EstimateS (Chapter 6, Colwell and Elsensohn 2014).

In Chapter 7 we explicitly mapped uncertainty resulting from picking one predictor out of a set of highly correlated predictors when predicting the current and forecasting the future distribution of introduced seaweeds. These results were then aggregated into change and turnover maps for which the standard deviation was mapped. However, species distribution modelling (SDM) is subject to multiple sources of uncertainty linked to the data, the model and the prediction (Rocchini et al., 2011; Beale & Lennon, 2012). Dormann et al. (2008) and Buisson et al. (2010) found that the SDM method contributes the most to the variation in future climate forecasts of species distribution. Other important factors include data uncertainty and the general circulation model (GCM) used for climate change impact studies (Dormann et al., 2008; Buisson et al., 2010). Synes and Osborne (2011), on the other hand, found that the predictor set used, contributes more to the uncertainty than the GCM and the climate change scenario. For maps of climate change forecasts of species distributions, prediction uncertainty assessment using residual variation analysis (PURV) plots allow for the visualization of interpolation and extrapolation uncertainties (Engler & Rodder, 2012). Concerning the uncertainty in the data, Naimi et al. (2011, 2014) showed that the impact of the positional uncertainty is heavily linked with the amount of spatial autocorrelation in the predictors. In order to visualize the uncertainty of species distribution maps, Rocchini et al. (2011) proposed to create maps of ignorance and Hartley et al. (2006) calculated confidence intervals for each location on the map.

The addition of an indication of the uncertainty in the OBIS quality control flags (Chapter 2), and more specifically the outlier flags, could be relevant for future studies. This could be achieved by distributing, next to the quality control flags themselves, the values that were used to calculate the outlier flags. Uncertainty in

the models due to errors in the environmental data could possibly be assessed by including predictor permutation procedures, based on the original data and an error model, in the R package *sdmpredictors* (Chapter 3). Bio-ORACLE (Tyberghein et al., 2012), for example, included maps of extrapolation errors which can be used as an error model. For other predictors, such as sea surface temperature, that didn't require extrapolation the error is uncertain. Generally, we expect these permutations to have the largest effect on the model at the edges of the species' distribution. But the impact on geographic space will depend on the interplay between the predictor importance, the spatial distribution of the errors and the distribution of the species modelled.

Another important aspect of SDM is the need to take into account the bias-variance trade off as the model complexity had a clear impact on our results. While we did not experiment with the parameters regulating the complexity of the algorithms, the impact of using algorithms with different complexities resulted in marked differences in the predictor relevance (Chapter 4). This was especially the case for generalized linear models (GLM) fitted with only linear features for which the sea surface temperature had a lower relevance in comparison to the other algorithms used. While modelling invasive seaweeds (Chapter 7) the complexity of the models was a priori restricted by only allowing quadratic and linear features in MaxEnt and GLM. Further restricting the model complexity lead to a decrease in model transferability.

The link between model complexity and the interpolation and extrapolation capabilities of species distribution models is well researched (Reineking & Schröder, 2006; Heikkinen et al., 2012; Syfert et al., 2013; Merow et al., 2014; Radosavljevic & Anderson, 2014; García-Callejas & Araújo, 2015; Moreno-Amat et al., 2015). Avoiding models which overfit the distribution only depends on the selection of the evaluation data which should penalize overly complex models. The validation data can be obtained from independent sampling campaigns or by subsampling the collected distribution records. Many researchers noted that when subsampling distribution records the spatio-temporal nature of the data should be taken into account by either using geographically or temporally independent data for validation (Fielding & Haworth, 1995; Boyce et al., 2002; Araujo et al., 2005; Veloz, 2009; Arlot & Celisse, 2010; Peterson et al., 2011; Anderson, 2012; Hijmans, 2012; Barbet-Massin & Jetz, 2014; Roberts et al., 2016). For invasive species with well-defined native ranges this is easy to perform as the native and invasive populations are spatially separated (Chapter 7, Jiménez-Valverde et al. 2011, Verbruggen et al. 2013, Petitpierre et al. 2017). For other species spatial data splitting is less obvious. In

Chapter 4 we proposed and implemented two spatial cross-validation methods: a disc-based and a grid-based approach. The biggest disadvantage of spatial cross-validation is that it can result in the need for extrapolation between the training and test set. This is desired when extrapolation is expected, which is the case for invasive species exhibiting considerable niche expansion or for climate change studies (Chapter 7). But, it can also lead to the selection of sub-optimal models (Anderson, 2012). This can in turn be mitigated by performing model selection based on the results of k-fold cross-validation (Roberts et al., 2016). In our predictor relevance study (Chapter 4) we noticed a marked difference between the relevance obtained by random and spatial cross-validation. These differences indicate that some predictors are more relevant for interpolation than for extrapolation.

Next to the use of virtual species (Box 2), the use of a benchmark dataset such as, MarineSPEED, is ideally suited for a further study of optimal model complexities and extrapolation abilities of marine species distribution models. This could be achieved by applying hyperparameter optimization techniques commonly used in machine learning in order to choose algorithm settings resulting in models that do not over- or underfit the species distribution. Care should hereby be taken to prioritize the ecological relevance and plausibility of the resulting models over the accuracy of the prediction as measured by the evaluation metric used (Bell & Schlaepfer, 2016; Brewer et al., 2016).

Some future perspectives

In this final section we would like to bring forward some general and specific recommendations for future research avenues related to modelling marine species distributions.

In Box 1 we already alluded to the possible added value of adding data such as traits data to taxonomic databases and biological and environmental data to biogeographic databases. Especially with more data on biotic interactions, migration patterns and environmental data becoming available, the ability to model the future distribution of marine species will greatly improve. Furthermore, an increased effort, from researchers, funding agencies and citizen science initiatives, in making the known distribution data publicly available will automatically lead towards better models as usually more data beats better algorithms (Halevy et al., 2009). More data could even lead to the development of general ecosystem models (GEM) with a higher resolution and precision as data is the main limiting factor for GEMs (Purves et al., 2013).

However, more data and more specifically more environmental data to choose from requires, next to a selection based on the ecology of a species, an easy way to filter out the most relevant predictors. For our predictor relevance study (Chapter 4) we unsuccessfully tried to predict the predictor ranking based on some rudimentary species traits and the ranking of models built with a single predictor. We believe that finding heuristics that are able to trim down or rank the extensive list of possible predictors of species distributions would be a valuable research topic. One possible avenue would be to elaborate on some of our preliminary results, based on a suggestion from Petitpierre et al. (2017), where we compared the performance of 10 random predictor sets and a set of 5 PCA variables on the MarineSPEED dataset and found that the best predictor set of the first fold of all combinations of predictors did not statistically have a better performance than the best predictor set out of 10 random predictor sets when tested on the other folds of a spatial cross-validation (Fig. 1). Another approach could involve modelling with all predictors and then, based on the importance of the variables in the model, reduce the number of included predictors as is done for Random Forests models in the R package *VSURF* (Genuer et al., 2015).

A combination of the automated selection of near-optimal predictors with additional research on the selection of algorithms and their parameters, can lead to the automated modelling of species distributions. Such a system would allow for numerous applications. As a first application, the generated models could be used as an additional quality control procedure in OBIS (Chapter 2) whereby records with a low predicted suitability would be flagged. Another use case could be the generation of climate change risk zones , especially if models from multiple species are combined into a community model and additional information on biotic interactions and the migratory capabilities of the species are included. As most marine invasive species are difficult to eradicate once established, there is a need for the monitoring of vulnerable habitats (Reiss et al., 2015). The generated species distribution models of invasive species could prove to be invaluable for this. But, additional information, such as the introduction vectors (shipping, aquaculture, …), will have to be taken into account.

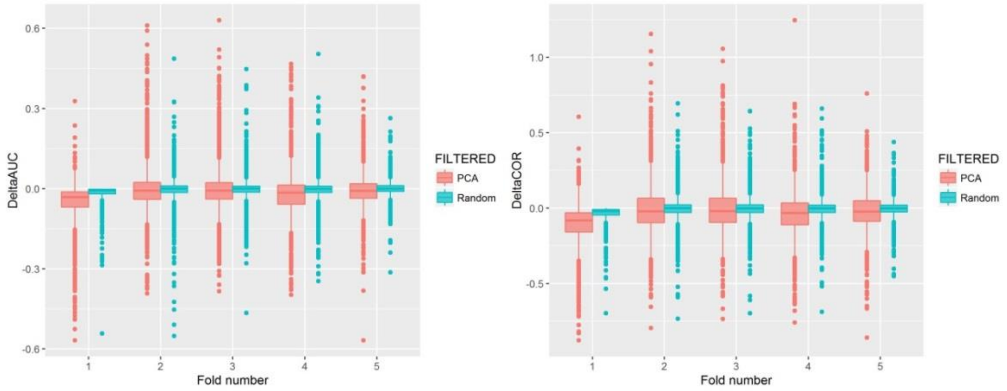


Figure 1. Performance of models build with PCA predictors or with random predictors measured as a delta the best predictor set from the first fold. Difference between the AUC (left) and COR (right) of the best predictor set from all combinations of 7 predictors from the first fold and the best predictor set from a set of 10 pseudo-random predictor sets from the first fold or the best model from a set of PCA models with different number of principal components from the first fold.

While performing niche analyses of invasive seaweeds in Chapter 7 with the intend to estimate the amount of niche expansion, several issues came up. Firstly, we noted that spatial thinning of distribution records, the study area and whether occurrences in novel environments are considered, all have an impact on the resulting niche metrics. While these changes have a moderate to high impact on some species, further study is needed to elucidate this by doing an in depth investigation of the generated kernel densities of multiple real and simulated invasive species. Secondly, the great discrepancy between calculating the niche expansion only for analog conditions or additionally including expansion into non-analog climates has to be further investigated. The biggest question therein is whether this niche expansion reflects genuine niche shifts with ecological or evolutionary mechanisms or whether it is due to methodological or data issues such as a lack of distribution records in the native habitat and differences in the available environmental conditions in the native and invaded area. Including this observed niche expansion either by transferring magnitude and character of observed niche shifts from well-studied avatar invaders to new or potential invaders or by the use of mechanistic models combining eco-physiological data with distribution data or some other method will no doubt lead to improved distribution models (Kearney & Porter, 2009; Larson et al., 2014). This is not only important for invasive species but will increasingly become important for the prediction of the future distribution of species under a changing global climate.

Box 2. Virtual species

Many researchers have used virtual species, sometimes in combination with real species, to study various aspects of species distribution modelling (SDM). The idea of virtual species is to simulate the species' probability of occurrence in relationship with one or more environmental gradients, and project it into a real or artificial landscape (Leroy et al., 2016).

Hirzel et al. (2001) pioneered this approach in their comparison on the performance of two methods (GLM and ENFA) for three different scenarios: a spreading species, a species at equilibrium and an overabundant species. Other studies used it to compare sample selection mitigation methods (Dudík et al., 2005; Kramer-Schadt et al., 2013; Varela et al., 2014), the performance of analysts (Austin et al., 2006), regularization methods (Reineking & Schröder, 2006), SDM algorithms (Real et al., 2006; Meynard & Quinn, 2007), methods accounting for spatial autocorrelation and collinearity (Dormann et al., 2007, 2013) and pseudo-absence generation methods (Wisz & Guisan, 2009). Furthermore virtual species have been used to assess the impact of data errors (Naimi et al., 2011, 2014; Lauzeral et al., 2012), species invasion (Václavík & Meentemeyer, 2009), scale effects (Bombi & D'Amen, 2012) and model complexity (Santika & Hutchinson, 2009; García-Callejas & Araújo, 2015) and to assess the general performance of SDM (Bahn & McGill, 2007; Elith & Graham, 2009; Zurell et al., 2009; Saupe et al., 2012; Li & Guo, 2013).

While most papers developed their own approach for creating virtual species, various software packages for virtual species have been (recently) released: COMPAS (Austin et al., 2006), demoniche (Nenzén et al., 2012), SDMvspecies (Duan et al., 2015), virtualspecies (Leroy et al., 2016), NicheLim (Huang et al., 2016) and NicheA (Qiao et al., 2016). Additionally Garzon-Lopez et al. (2016) released a set of two virtual terrestrial species. Some commonly used approaches for simulating species include drawing a distribution from a known equation in a real or simulated environment, sampling a spatially explicit population model or creating a thresholded species distribution model and then subsampling from the resulting map (Miller, 2014).

As the true distribution of the virtual species is known, this approach allows for a direct comparison with the predictions from an SDM. The use of virtual species thus allows for the independent testing of various properties affecting the performance of species distribution models without the presence of any confounding factors, such as sample selection bias, commonly present in real datasets (Miller, 2014). Virtual species are particularly useful if relevant aspects of the species distribution, such as population processes and competition are included (Elith & Graham, 2009). The biggest disadvantage of using virtual species is that the ecological relevance and transferability of the obtained results towards real species is often not assessed (Meynard & Kaplan, 2013; Miller, 2014). A mixed approach, combining virtual and real species such as those from the MarineSPEED benchmark dataset (Chapter 4) could alleviate this concern.

References

Acevedo P., Jiménez-Valverde A., Lobo J.M., & Real R. (2012) Delimiting the geographical background in species distribution modelling. *Journal of Biogeography*, **39**, 1383–1390.

Acuna E. & Rodriguez C. (2004) A meta analysis study of outlier detection methods in classification. *Proceedings of the International IPSI 2004 Conference, Symposium on Challenges in Internet and Interdisciplinary Research*.

Adams N.M. (1994) *Seaweeds of New Zealand. An Illustrated Guide.* Canterbury University Press, Christchurch.

Adey W.H. & Steneck R.S. (2001) Thermogeography over time creates biogeographic regions: A temperature/space/time-integrated model and an abundance-weighted test for benthic marine algae. *Journal of Phycology*, **37**, 677–698.

Aiello-Lammens M.E., Boria R. a, Radosavljevic A., Vilela B., & Anderson R.P. (2015) spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, **38**, 541–545.

Aleem A.A. (1948) The recent migration of certain Indopacific algae from the Red Sea into the Mediterranean. *New Phytologist*, **47**, 88–94.

Aleem A.A. (1950) Some new records of marine algae from the Mediterranean Sea with reference to their geographical distribution. *Acta Horti Gothoburgensis*, **18**, 276–288.

Allouche O., Tsoar A., & Kadmon R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

Amante C. & Eakins B.. (2009) ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Tech. Memo. NESDIS NGDC-24.

Anderson R.P. (2012) Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. *Annals of the New York Academy of Sciences*, **1260**, 66–80.

Anderson R.P. & Gonzalez I. (2011) Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent. *Ecological Modelling*, **222**, 2796–2811.

Anderson R.P. & Raza A. (2010) The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: Preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *Journal of Biogeography*, **37**, 1378–1393.

Andreakis N., Procaccini G., Maggs C.A., & Kooistra W.H.C.F. (2007) Phylogeography of the invasive seaweed *Asparagopsis* (Bonnemaisoniales, Rhodophyta) reveals cryptic diversity. *Molecular Ecology*, **16**, 2285–2299.

Andreakis N. & Schaffelke B. (2012) Invasive Marine Seaweeds: Pest or Prize? *Seaweed biology: Novel Insights into Ecophysiology, Ecology and Utilization* pp. 235–262. Springer Berlin Heidelberg, Berlin, Heidelberg.

Anon. (2010) *General Bathymetric Chart of the Oceans (GEBCO): the GEBCO_08 Grid.* General Bathymetric Chart of the Oceans, British Oceanographic Data Centre (BODC).

Appeltans W., Ahyong S.T., Anderson G. *et al.* (2012) The Magnitude of Global Marine Species Diversity. *Current Biology*, **22**, 2189–2202.

Aragay J., Vitales D., Gómez Garreta A., Ribera Siguan M.A., Steen F., De Clerck O., Garnatje T., & Rull Lluch J. (2016) Phenological and molecular studies on the introduced seaweed *Dictyota cyanoloma* (Dictyotales, Phaeophyceae) along the Mediterranean coast of the Iberian Peninsula. *Mediterranean Marine Science*, **17**, 766–776.

Araujo M.B., Pearson R.G., Thuiller W., & Erhard M. (2005) Validation of species-climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Araújo M.B. & Guisan A. (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33**, 1677–1688.

Araújo M.B. & New M. (2007) Ensemble forecasting of species distributions. *Trends in ecology & evolution*, **22**, 42–7.

Araújo M. & Peterson A. (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology*, **93**, 1527–1539.

Arlot S. & Celisse A. (2010) A survey of cross-validation procedures for model selection. *Statistics Surveys*, **4**, 40–79.

Assis J., Lucas A.V., Bárbara I., & Serrão E.Á. (2014) Future climate change is predicted to shift long-term persistence zones in the cold-temperate kelp Laminaria hyperborea. *Marine Environmental Research*, **113**, 174–182.

Assis J., Zupan M., Nicastro K.R., Zardi G.I., McQuaid C.D., & Serrão E.A. (2015) Oceanographic Conditions Limit the Spread of a Marine Invader along Southern African Shores. *PLOS ONE*, **10**, e0128124.

Assis J., Coelho N.C., Lamy T., Valero M., Alberto F., & Serrão E.A. (2016) Deep reefs are climatic refugia for genetic diversity of marine forests. *Journal of Biogeography*, 833–844.

August T., Lucas T., Golding N., van Loon E., & Mcinerny G. (2017) *zoon: Reproducible, Accessible & Shareable Species Distribution Modelling*. R package version 0.6. Available at: http://cran.r-project.org/package=zoon.

Austin M.P. (1980) Searching for a Model for Use in Vegetation Analysis. *Classification and Ordination* pp. 11–21. Springer Netherlands, Dordrecht.

Austin M.P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101–118.

Austin M.P., Belbin L., Meyers J. a., Doherty M.D., & Luoto M. (2006) Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. *Ecological Modelling*, **199**, 197–216.

Bahn V. & McGill B.J. (2007) Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, **16**, 733–742.

Banks W.E., D'Errico F., Peterson A.T., Vanhaeren M., Kageyama M., Sepulchre P., Ramstein G., Jost A., & Lunt D. (2008) Human ecological niches and ranges during the LGM in Europe derived from an application of eco-cultural niche modeling. *Journal of Archaeological Science*, **35**, 481–491.

Barbet-Massin M., Jiguet F., Albert C.H., & Thuiller W. (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.

Barbet-Massin M. & Jetz W. (2014) A 40-year, continent-wide, multispecies assessment of relevant climate predictors for species distribution modelling. *Diversity and Distributions*, **20**, 1285–1295.

Barnes M.A., Jerde C.L., Wittmann M.E., Chadderton W.L., Ding J., Zhang J., Purcell M., Budhathoki M., & Lodge D.M. (2014) Geographic selection bias of occurrence data influences transferability of invasive *Hydrilla verticillata* distribution models. *Ecology and Evolution*, **4**, 2584–2593.

Barry S. & Elith J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413–423.

Barve N., Barve V., Jiménez-Valverde A., Lira-Noriega A., Maher S.P., Peterson A.T., Soberón J., & Villalobos F. (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, **222**, 1810–1819.

Bates A.E., Bird T.J., Stuart-Smith R.D., Wernberg T., Sunday J.M., Barrett N.S., Edgar G.J., Frusher S., Hobday A.J., Pecl G.T., Smale D.A., & McCarthy M. (2015) Distinguishing geographical range shifts from artefacts of detectability and sampling effort. *Diversity and Distributions*, **21**, 13–22.

Beale C.M. & Lennon J.J. (2012) Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 247–258.

Beaugrand G., Lenoir S., Ibañez F., & Manté C. (2011) A new model to assess the probability of occurrence of a species, based on presence-only data. *Marine Ecology Progress Series*, **424**, 175–190.

Beck J., Böller M., Erhardt A., & Schwanghart W. (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, **19**, 10–15.

Beissbarth T., Fellenberg K., Brors B., Arribas-Prat R., Boer J., Hauser N.C., Scheideler M., Hoheisel J.D., Schütz G., Poustka A., & Vingron M. (2000) Processing and quality control of DNA array hybridization data. *Bioinformatics (Oxford, England)*, **16**, 1014–22.

Belanger C.L., Jablonski D., Roy K., Berke S.K., Krug A.Z., & Valentine J.W. (2012) Global environmental predictors of benthic marine biogeographic structure. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 14046–51.

Bell D.M. & Schlaepfer D.R. (2016) On the dangers of model complexity without ecological justification in species distribution modeling. *Ecological Modelling*, **330**, 50–59.

Belton G.S., van Reine W.F.P., Huisman J.M., Draisma S.G.A., & D. Gurgel C.F. (2014) Resolving phenotypic plasticity and species designation in the morphologically challenging *Caulerpa racemosa* - peltata complex (Chlorophyta, Caulerpaceae). *Journal of Phycology*, **50**, 32–54.

Benito B.M., Svenning J.-C., Kellberg-Nielsen T., Riede F., Gil-Romera G., Mailund T., Kjaergaard P.C., & Sandel B.S. (2017) The ecological niche and distribution of Neanderthals during the Last Interglacial. *Journal of Biogeography*, **44**, 51–61.

Bentlage B., Peterson A.T., Barve N., & Cartwright P. (2013) Plumbing the depths: extending ecological niche modelling and species distribution modelling in three dimensions. *Global Ecology and Biogeography*, **22**, 952–961.

Bianchi, Carlo N. & Morri C. (2003) Global sea warming and "tropicalization" of the Mediterranean Sea: biogeographic and ecological aspects. *Biogeographia – The Journal of Integrative Biogeography*, **24**, 319–329.

Bingham H., Doudin M., Weatherdon L. *et al.* (2017) The Biodiversity Informatics Landscape: Elements, Connections and Opportunities. *Research Ideas and Outcomes*, **3**, e14059.

Boavida J., Assis J., Silva I., & Serrão E.A. (2016) Overlooked habitat of a vulnerable gorgonian revealed in the Mediterranean and Eastern Atlantic by ecological niche modelling. *Scientific Reports*, **6**, 36460.

Bocedi G., Palmer S.C.F., Pe'er G., Heikkinen R.K., Matsinos Y.G., Watts K., & Travis J.M.J. (2014) RangeShifter: a platform for modelling spatial eco-evolutionary dynamics and species' responses to environmental changes. *Methods in Ecology and Evolution*, **5**, 388–396.

Bolton T.F. & Graham W.M. (2006) Jellyfish on the Rocks: Bioinvasion Threat of the International Trade in Aquarium Live Rock. *Biological Invasions*, **8**, 651–653.

Bombi P. & D'Amen M. (2012) Scaling down distribution maps from atlas data: a test of different approaches with virtual species. *Journal of Biogeography*, **39**, 640–651.

Booth T.H., Nix H.A., Busby J.R., & Hutchinson M.F. (2014) BIOCLIM : the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions*, **20**, 1–9.

Boria R.A., Olson L.E., Goodman S.M., & Anderson R.P. (2014) Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, **275**, 73–77.

Bosch S., Tyberghein L., & De Clerck O. (2016) *sdmpredictors: Species Distribution Modelling Predictor Datasets*. R package version 0.9. Available at: https://github.com/lifewatch/sdmpredictors.

Bosch S., Tyberghein L., Deneudt K., Hernandez F., & De Clerck O. (2017) *marinespeed: Benchmark Data Sets and Functions for Marine Species Distribution Modelling*. R package version 0.1.0. Available at: https://cran.r-project.org/package=marinespeed.

Boudouresque C.F. (1999) The Red Sea-Mediterranean link: unwanted effects of canals. *Invasive species and biodiversity management* pp. 213–228. Kluwer Academic Publishers, Dordrecht, NL.

Boudouresque C.F. & Verlaque M. (2002) Biological pollution in the Mediterranean Sea: invasive versus introduced macrophytes. *Marine Pollution Bulletin*, **44**, 32–38.

Boudouresque C.F., Klein J., Ruitton S., & Verlaque M. (2010) Biological Invasion: The Thau Lagoon, a Japanese Biological Island in the Mediterranean Sea. *Global Change: Mankind-Marine Environment Interactions* pp. 151–156. Springer Netherlands, Dordrecht.

Boyce M.S., Vernier P.R., Nielsen S.E., & Schmiegelow F.K.. (2002) Evaluating resource selection functions. *Ecological Modelling*, **157**, 281–300.

Bradie J. & Leung B. (2016) A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. *Journal of Biogeography*.

Brandt L.A., Benscoter A.M., Harvey R., Speroterra C., Bucklin D., Romañach S.S., Watling J.I., & Mazzotti F.J. (2017) Comparison of climate envelope models developed using expert-selected variables versus statistical selection. *Ecological Modelling*, **345**, 10–20.

Braunisch V., Coppes J., Arlettaz R., Suchant R., Schmid H., & Bollmann K. (2013) Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change. *Ecography*, **36**, 971–983.

Breeman A.M. (1988) Relative importance of temperature and other factors in determining geographic boundaries of seaweeds: Experimental and phenological evidence. *Helgolander Meeresuntersnder Meeresuntersuchungen*, **42**, 199–241.

Breiman L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Brewer M.J., O'Hara R.B., Anderson B.J., & Ohlemüller R. (2016) Plateau : a new method for ecologically plausible climate envelopes for species distribution modelling. *Methods in Ecology and Evolution*, **7**, 1489–1502.

Briones C., Rivadeneira M.M., Fernández M., & Guiñez R. (2014) Geographical Variation of Shell Thickness in the Mussel *Perumytilus purpuratus* Along the Southeast Pacific Coast. *The Biological Bulletin*, **227**, 221–231.

Brodie J., Bartsch I., Neefus C., Orfanidis S., Bray T., & Mathieson A.C. (2007a) New insights into the cryptic diversity of the North Atlantic–Mediterranean " *Porphyra leucosticta* " complex: *P. olivii* sp. nov. and *P. rosengurttii* (Bangiales, Rhodophyta). *European Journal of Phycology*, **42**, 3–28.

Brodie J., Maggs C.A., John D.M., & Blomster J. (2007b) *Green seaweeds of Britain and Ireland.* British Phycological Society, London, UK.

Broennimann O. & Guisan A. (2008) Predicting current and future biological invasions: both native and invaded ranges matter. *Biology letters*, **4**, 585–9.

Broennimann O., Fitzpatrick M.C., Pearman P.B., Petitpierre B., Pellissier L., Yoccoz N.G., Thuiller W., Fortin M.-J., Randin C., Zimmermann N.E., Graham C.H., & Guisan A. (2012) Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography*, **21**, 481–497.

Broennimann O., Petitpierre B., Randin C. *et al.* (2016) *ecospat: Spatial Ecology Miscellaneous Methods*. R package version 1.1. Available at: http://cran.r-project.org/package= ecospat.

Brook B.W., Sodhi N.S., & Bradshaw C.J.A. (2008) Synergies among extinction drivers under global change. *Trends in Ecology and Evolution*, **23**, 453–460.

Bucklin D.N., Basille M., Benscoter A.M., Brandt L.A., Mazzotti F.J., Romañach S.S., Speroterra C., & Watling J.I. (2015) Comparing species distribution models constructed with different subsets of environmental predictors. *Diversity and Distributions*, **21**, 23–35.

Buisson L., Thuiller W., Casajus N., Lek S., & Grenouillet G. (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, **16**, 1145–1157.

Busby J.R. (1991) BIOCLIM-a bioclimate analysis and prediction system. *Nature Conservation: Cost Effective Biological Surveys and Data Analysis* (ed. by C.R. Margules and M.P. Austin), pp. 64–68. CSIRO, Canberra.

Caldow C., Monaco M.E., Pittman S.J., Kendall M.S., Goedeke T.L., Menza C., Kinlan B.P., & Costa B.M. (2015) Biogeographic assessments: A framework for information synthesis in marine spatial planning. *Marine Policy*, **51**, 423–432.

Candela L., Castelli D., Coro G., Lelii L., Mangiacrapa F., Marioli V., & Pagano P. (2015) An infrastructure-oriented approach for supporting biodiversity research. *Ecological Informatics*, **26**, 162–172.

de Candolle A. (1855) *Géographie botanique raisonnée; ou, Exposition des faits principaux et des lois concernant la distribution géographique des plantes de l'époque actuelle.* V. Masson, Paris, FR.

Carpenter G., Gillison A.N., & Winter J. (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, **2**, 667–680.

Cecere E., Alabiso G., Carlucci R., Petrocelli A., & Verlaque M. (2016) Fate of two invasive or potentially invasive alien seaweeds in a central Mediterranean transitional water system: failure and success. *Botanica Marina*, **59**, 451-462.

Chamberlain S., Boettiger C., Karthik R., Barve V., & Mcglinn D. (2016a) *rgbif: Interface to the Global Biodiversity Information Facility API*. R package version 0.9.3. Available at: https://github.com/ropensci/rgbif.

Chamberlain S., Michonneau F., Provoost P., & Sumner M. (2016b) *mregions: Marine Regions Data from "Marineregions.org."* R package version 0.1.4. Available at: https://github.com/ropenscilabs/mregions.

Chandler M., See L., Copas K., Bonde A.M.Z., López B.C., Danielsen F., Legind J.K., Masinde S., Miller-Rushing A.J., Newman G., Rosemartin A., & Turak E. (2016) Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*.

Chang W., Cheng J., Allaire J., Xie Y., & McPherson J. (2016) *shiny: Web Application Framework for R*. R package version 0.14.1. Available at: http://shiny.rstudio.com.

Chapman A. (2005) *Principles of data quality, version 1.0, Report for the Global Biodiversity Information Facility.* GBIF Secretariat, Copenhagen.

Charney N.D. (2012) Evaluating expert opinion and spatial scale in an amphibian model. *Ecological Modelling*, **242**, 37–45.

Chaudhary C., Saeedi H., & Costello M.J. (2016) Bimodality of Latitudinal Gradients in Marine Species Richness. *Trends in Ecology & Evolution*, **31**, 670–676.

Chazdon R.L., Colwell R.K., Denslow J.S., & Guariguata M.R. (1998) Statistical methods for estimating species richness of woody regeneration in primary and secondary rainforests of northeastern Costa Rica. *Forest biodiversity research, monitoring and modeling: Conceptual background and Old World case studies* pp. 285–309. Parthenon Publishing, Paris, France.

Chefaoui R.M. & Lobo J.M. (2008) Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, **210**, 478–486.

Chen I.-C., Hill J.K., Ohlemuller R., Roy D.B., & Thomas C.D. (2011) Rapid Range Shifts of Species Associated with High Levels of Climate Warming. *Science*, **333**, 1024–1026.

Cheung W.W.L., Lam V.W.Y., & Pauly D. (2008) Dynamic bioclimate envelope model to predict climate-induced changes in distribution of marine fishes and invertebrates. *Fisheries Centre Research Report 16*, **16**, 5–50.

Chiarucci A., Bacaro G., & Scheiner S.M. (2011) Old and new challenges in using species diversity for assessing biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**, 2426–2437.

Chualáin F.N., Maggs C.A., Saunders G.W., & Guiry M.D. (2004) The invasive genus *Asparagopsis* (Bonnemaisoniaceae, Rhodophyta): Molecular Systematics, Morphology, and Ecophysiology of Falkenbergia isolates. *Journal of Phycology*, **40**, 1112–1126.

CITES (2006) Available at: https://cites.org/sites/default/files/common/com/ac/22/E22i-08.pdf.

Clark J.D., Dunn J.E., & Smith K.G. (1993) A Multivariate Model of Female Black Bear Habitat Use for a Geographic Information System. *The Journal of Wildlife Management*, **57**, 519.

Clark J.S., Gelfand A.E., Woodall C.W., & Zhu K. (2014) More than the sum of the parts: Forest Climate response from join specise distributions. *Ecological Applications*, **24**, 990–999.

Claus S., De Hauwere N., Vanhoorne B., Deckers P., Souza Dias F., Hernandez F., & Mees J. (2014) Marine Regions: Towards a Global Standard for Georeferenced Marine Names and Boundaries. *Marine Geodesy*, **37**, 99–125.

Coll M., Piroddi C., Steenbeek J. *et al.* (2010) The Biodiversity of the Mediterranean Sea: Estimates, Patterns, and Threats. *PLoS ONE*, **5**, e11842.

Colwell R.K. & Elsensohn J.E. (2014) EstimateS turns 20: Statistical estimation of species richness and shared species from samples, with non-parametric extrapolation. *Ecography*, **37**, 609–613.

Colwell R.K. & Rangel T.F. (2009) Hutchinson's duality: The once and future niche. *Proceedings of the National Academy of Sciences*, **106**, 19651–19658.

Combal B. (2014) Available at: http://doi.org/10.5281/zenodo.12781.

Congalton R.G. (1991) A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, **37**, 35–46.

Convey P., Chown S.L., Clarke A. *et al.* (2014) The spatial structure of Antarctic biodiversity. *Ecological Monographs*, **84**, 203–244.

Coro G., Magliozzi C., Vanden Berghe E., Bailly N., Ellenbroek A., & Pagano P. (2016) Estimating absence locations of marine species from data of scientific surveys in OBIS. *Ecological Modelling*, **323**, 61–76.

Coro G., Webb T.J., Appeltans W., Bailly N., Cattrijsse A., & Pagano P. (2015) Classifying degrees of species commonness: North Sea fish as a case study. *Ecological Modelling*, **312**, 272–280.

Corriero G., Pierri C., Accoroni S. *et al.* (2016) Ecosystem vulnerability to alien and invasive species: a case study on marine habitats along the Italian coast. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **26**, 392–409.

Cortes C. & Vapnik V. (1995) Support-vector networks. *Machine Learning*, **20**, 273–297.

Costa H., Ponte N.B., Azevedo E.B., & Gil A. (2015) Fuzzy set theory for predicting the potential distribution and cost-effective monitoring of invasive species. *Ecological Modelling*, **316**, 122–132.

Costello M.J., Michener W.K., Gahegan M., Zhang Z.-Q., & Bourne P.E. (2013) Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution*, **28**, 454–461.

Crimmins S.M., Dobrowski S.Z., & Mynsberge A.R. (2013) Evaluating ensemble forecasts of plant species distributions under climate change. *Ecological Modelling*, **266**, 126–130.

Cucuringu M. (2016) Sync-Rank: Robust Ranking, Constrained Ranking and Rank Aggregation via Eigenvector and SDP Synchronization. *IEEE Transactions on Network Science and Engineering*, **3**, 58–79.

Dambach J. & Rödder D. (2011) Applications and future challenges in marine species distribution modeling. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **21**, 92–100.

Davies A.J. & Guinotte J.M. (2011) Global Habitat Suitability for Framework-Forming Cold-Water Corals. *PLOS ONE*, **6**, 1–15.

Davies L. & Gather U. (1993) The Identification of Multiple Outliers. *Journal of the American Statistical Association*, **88**, 782–792.

DeWalt S.J., Denslow J.S., & Ickes K. (2004) Natural-enemy release facilitates habitat expansion of the invasive tropical shrub Clidemia hirta. *Ecology*, **85**, 471–483.

Doelle M., McConnell M.L., & VanderZwaag D.L. (2007) Invasive seaweeds: global and regional law and policy responses. *Botanica Marina*, **50**, 438–450.

Doney S.C., Ruckelshaus M., Emmett Duffy J. *et al.* (2012) Climate Change Impacts on Marine Ecosystems. *Annual Review of Marine Science*, **4**, 11–37.

Dormann C.F., McPherson J.M., Araújo M.B. *et al.* (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**, 609–628.

Dormann C.F., Purschke O., García Márquez J.R., Lautenbach S., & Schröder B. (2008) Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. *Ecology*, **89**, 3371–86.

Dormann C.F., Elith J., Bacher S. *et al.* (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 027–046.

Downie A.-L., von Numers M., & Boström C. (2013) Influence of model selection on the predicted distribution of the seagrass *Zostera marina*. *Estuarine, Coastal and Shelf Science*, **121–122**, 8–19.

Duan R., Kong X., Huang M., Wu G., & Wang Z. (2015) SDMvspecies : a software for creating virtual species for species distribution modelling. *Ecography*, **38**, 108–110.

Dudík M., Schapire R.E., & Phillips S.J. (2005) Correcting sample selection bias in maximum entropy density estimation. *Advances in Neural Information Processing Systems*, **18**, 323–330.

Duque-Lazo J., van Gils H., Groen T. a., & Navarro-Cerrillo R.M. (2016) Transferability of species distribution models: The case of *Phytophthora cinnamomi* in Southwest Spain and Southwest Australia. *Ecological Modelling*, **320**, 62–70.

Early R. & Sax D.F. (2014) Climatic niche shifts between species' native and naturalized ranges raise concern for ecological forecasts during invasions and climate change. *Global Ecology and Biogeography*, **23**, 1356–1365.

Edgar G. & Stuart-Smith R. (2009) Ecological effects of marine protected areas on rocky reef communities—a continental-scale analysis. *Marine Ecology Progress Series*, **388**, 51–62.

Edgar G.J., Bates A.E., Bird T.J., Jones A.H., Kininmonth S., Stuart-Smith R.D., & Webb T.J. (2016) New Approaches to Marine Conservation Through Scaling Up of Ecological Data. *Annual Review of Marine Science*, **8**, 435-461.

Efron B. & Tibshirani R. (1986) Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, **1**, 54–75.

Eggert A. (2012) Seaweed Responses to Temperature. *Seaweed biology: Novel Insights into Ecophysiology, Ecology and Utilization* (ed. by C. Wiencke and K. Bischof), pp. 47–66. Springer, Berlin, Heidelberg.

Elith J., Graham C.H., Anderson R.P. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith J. & Graham C.H. (2009) Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, **32**, 66–77.

Elith J. & Leathwick J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Elith J., Kearney M., & Phillips S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Elith J., Phillips S.J., Hastie T., Dudík M., Chee Y.E., & Yates C.J. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Elton C. (1927) *Animal Ecology.* Sedgwick and Jackson, London, UK.

Engler J.O. & Rodder D. (2012) Disentangling interpolation and extrapolation uncertainties in ecologial niche models : a novel visualization technique for the spatial variation of predictor variable colinearity. *Biodiversity Informatics*, **8**, 30–40.

European Commission (2016) Commission Implementing Regulation (EU) 2016/1141 of 13 July 2016 adopting a list of invasive alien species of Union concern pursuant to Regulation (EU) No 1143/2014 of the European Parliament and of the Council. *Official Journal of the European Union*.

European Parliament (2014) Regulation (EU) No 1143/2014 of the European Parliament and of the Council of 22 October 2014 on the prevention and management of the introduction and spread of invasive alien species. *Official Journal of the European Union*.

Falls W.W., Ehringer J.N., Herndon R., Herndon T., Nichols M., Nettles S., Armstrong C., & Haverkamp D. (2008) Aquacultured Live Rock as an Alternative to Imported Wild-Harvested Live Rock: An Update. *Marine Ornamental Species* pp. 207–218. Blackwell Publishing Company, Ames, Iowa, USA.

Fernández D. & Nakamura M. (2015) Estimation of spatial sampling effort based on presence-only data and accessibility. *Ecological Modelling*, **299**, 147–155.

Fielding A.H. & Haworth P.F. (1995) Testing the Generality of Bird-Habitat Models. *Conservation Biology*, **9**, 1466–1481.

Flagella M.M., Verlaque M., Soria A., & Buia M.C. (2007) Macroalgal survival in ballast water tanks. *Marine Pollution Bulletin*, **54**, 1395–1401.

Fogarty H.E., Burrows M.T., Pecl G.T., Robinson L.M., & Poloczanska E.S. (2017) Are fish outside their usual ranges early indicators of climate-driven range shifts? *Global Change Biology*.

Fogel F., D'Aspremont A., & Vojnovic M. (2016) Spectral Ranking using Seriation. *Journal of Machine Learning Research*, **17**, 1–45.

Fosså S. & Nilsen A. (1996) *The modern coral reef aquarium.* Birgit Schmettkamp Verlag, Berlin.

Fourcade Y., Engler J.O., Rödder D., & Secondi J. (2014) Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. *PLoS ONE*, **9**, 1–13.

Franklin J. (1995) Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography*, **19**, 474–499.

Franklin J. (2009) *Mapping species distributions. Spatial inference and prediction.* Cambridge University Press, Cambridge, USA.

Freeman E. a. & Moisen G.G. (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, **217**, 48–58.

Friedman J.H. (1991) Multivariate Adaptive Regression Splines. *The Annals of Statistics*, **19**, 1–67.

Friedman J.H. (2001) Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, **29**, 1189-1232.

Froese R., Bailly N., Coronado G.U., Pruvost P., Reyes R., & Hureau J.-C. (1999) A new procedure to clean up fish collection databases. *Proceedings of 5th Indo-Pacific Fish Conference* (eds Sire J-Y and Séret B). Society of French Ichthyologists, Paris, pp. 697–705.

Froese R., Pauly D., & Editors (2017) FishBase. Available from http://www.fishbase.org. Accessed 2017-05-11.

Galil B.S., Boero F., Campbell M.L. *et al.* (2015) "Double trouble": the expansion of the Suez Canal and marine bioinvasions in the Mediterranean Sea. *Biological Invasions*, **17**, 973–976.

Gallardo B., Clavero M., Sánchez M.I., & Vilà M. (2016a) Global ecological impacts of invasive species in aquatic ecosystems. *Global Change Biology*, **22**, 151–163.

Gallardo T., Bárbara I., Alfonso-Carrillo J., Bermejo R., Altamirano M., Gómez-Garreta A., Barceló C., Rull J., Ballesteros E., & De la Rosa J. (2016b) *Nueva lista crítica de las algas bentónicas marinas de España.* Sociedad Española de Ficología, Granada.

Garaba S.P., Wernand M.R., & Zielinski O. (2011) Quality control of automated hyperspectral remote sensing measurements from a seaborne platform. *Ocean Science Discussions*, **8**, 613–638.

García-Callejas D. & Araújo M.B. (2015) The effects of model and data complexity on predictions from species distributions models. *Ecological Modelling*, **326**, 4-12.

Garzon-Lopez C.X., Bastin L., Foody G.M., & Rocchini D. (2016) A virtual species set for robust and reproducible species distribution modelling tests. *Data in Brief*, **7**, 476–479.

GBIF (2011) *Darwin Core Archive Format, Reference Guide to the XML Descriptor File, April 2011, (contributed by Döring, M., Robertson, T., Remsen, D.).* Global Biodiversity Information Facility, Copenhagen.

Geijzendorffer I.R., Regan E.C., Pereira H.M. *et al.* (2016) Bridging the gap between biodiversity data and policy reporting needs: An Essential Biodiversity Variables perspective. *Journal of Applied Ecology*, **53**, 1341–1350.

Génard M. & Lescourret F. (2013) Combining niche and dispersal in a simple model (NDM) of species distribution. *PloS one*, **8**, e79948.

Genuer R., Poggi J.-M., & Tuleau-Malot C. (2015) VSURF: An R Package for Variable Selection Using Random Forests. *The R Journal*, **7**, 19–33.

Georges D. & Thuiller W. (2013) Multi-species distribution modeling with biomod2.

Gil-Rodríguez M.C., Haroun M.C., Ojeda Rodríguez A., Berecibar Zugasti E., & Domínguez Santana, P. Herrera Morán B. (2003) Proctoctista. *Lista de especies marinas de Canarias (algas, hongos, plantas y animales)* (ed. by L. Moro, J.L. Martín, M.J. Garrido, and I. Izquierdo), pp. 5–30. Consejería de Política Territorial y Medio Ambiente del Gobierno de Canarias, Las Palmas.

Golding N. & Purse B. V. (2016) Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, **7**, 598–608.

Gollasch S. (2006) Overview on introduced aquatic species in European navigational and adjacent waters. *Helgoland Marine Research*, **60**, 84–89.

Gollasch S. (2007) Is Ballast Water a Major Dispersal Mechanism for Marine Organisms? *Biological Invasions* (ed. by W. Nentwig), pp. 49–57. Springer Berlin Heidelberg, Berlin, Heidelberg.

Gotelli N. & Colwell R. (2010) Estimating species richness. *Biological Diversity. Frontiers in Measurement and Assessment*, 39–54.

Graham C.H., Elith J., Hijmans R.J., Guisan A., Peterson A.T., & Loiselle B.A. (2007a) The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, **45**, 239–247.

Graham M.H., Kinlan B.P., Druehl L.D., Garske L.E., & Banks S. (2007b) Deep-water kelp refugia as potential hotspots of tropical marine diversity and productivity. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 16576–16580.

Grassle F. (2000) The Ocean Biogeographic Information System (OBIS): An On-line, Worldwide Atlas for Accessing, Modeling and Mapping Marine Biological Data in a Multidimensional Geographic Context. *Oceanography*, **13**, 5–7.

Griffiths H.J. & Waller C.L. (2016) The first comprehensive description of the biodiversity and biogeography of Antarctic and Sub-Antarctic intertidal communities. *Journal of Biogeography*, **43**, 1143–1155.

Grinnell J. (1904) The origin and distribution of the chestnut-backed chickadee. *The Auk*, **21**, 364–382.

Grinnell J. (1917) The Niche-Relationships of the California Thrasher. *The Auk*, **34**, 427–433.

Guidetti P., Magnali L., & Navone A. (2015) First record of the acanthurid fish *Zebrasoma xanthurum* (Blyth, 1852) in the Mediterranean Sea, with some considerations on the risk associated with aquarium trade. *Mediterranean Marine Science*, **17**.

Guiry M.D. & Guiry G.M. (2017) Available at: http://www.algaebase.org.

Guisan A. & Theurillat J.-P. (2000) Equilibrium modeling of alpine plant distribution: how far can we go? *Phytocoenologia*, **30**, 353–384.

Guisan A. & Zimmermann N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Guisan A. & Thuiller W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Guisan A., Lehmann A., Ferrier S., Austin M., Overton J.M.C., Aspinall R., & Hastie T. (2006) Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology*, **43**, 386–392.

Guisan A., Graham C.H., Elith J. *et al.* (2007a) Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, **13**, 332–340.

Guisan A., Zimmermann N.E., Elith J., Graham C.H., Phillips S., & Peterson A.T. (2007b) What matters for predicting the occurrence of trees: Techniques, data, or species' characteristics? *Ecological Monographs*, **77**, 615–630.

Guisan A., Petitpierre B., Broennimann O., Kueffer C., Randin C., & Daehler C. (2012) Response to Comment on "Climatic Niche Shifts Are Rare Among Terrestrial Plant Invaders." *Science*, **338**, 193–193.

Guisan A., Petitpierre B., Broennimann O., Daehler C., & Kueffer C. (2014) Unifying niche shift studies: Insights from biological invasions. *Trends in Ecology and Evolution*, **29**, 260–269.

Gutt J., Zurell D., Bracegridle T. et al. (2012) Correlative and dynamic species distribution modelling for ecological predictions in the Antarctic: a cross-disciplinary concept. *Polar Research*, **31**.

Halevy A., Norvig P., & Pereira F. (2009) The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, **24**, 8–12.

Hamilton S.H., Pollino C. a., & Jakeman A.J. (2015) Habitat suitability modelling of rare species using Bayesian networks: Model evaluation under limited data. *Ecological Modelling*, **299**, 64–78.

Hammann M., Buchholz B., Karez R., & Weinberger F. (2013) Direct and indirect effects of Gracilaria vermiculophylla on native *Fucus vesiculosus*. *Aquatic Invasions*, **8**, 121–132.

Hanberry B.B., He H.S., & Palik B.J. (2012) Pseudoabsence generation strategies for species distribution models. *PloS one*, **7**, e44486.

Hand D.J. (2009) Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, **77**, 103–123.

Hanley J.A. & McNeil B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Harley C.D.G., Anderson K.M., Demes K.W., Jorve J.P., Kordas R.L., Coyle T.A., & Graham M.H. (2012) Effects of climate change on global seaweed communities. *Journal of Phycology*, **48**, 1064–1078.

Harley C.D.G., Hughes A.R., Hultgren K.M., Miner B.G., Sorte C.J.B., Thornber C.S., Rodriguez L.F., Tomanek L., & Williams S.L. (2006) The impacts of climate change in coastal marine systems. *Ecology Letters*, **9**, 228–241.

Harris D.J. (2015) Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, **6**, 465–473.

Hartley S., Harris R., & Lester P.J. (2006) Quantifying uncertainty in the potential distribution of an invasive species: climate and the Argentine ant. *Ecology letters*, **9**, 1068–79.

Hastie T. & Tibshirani R. (1986) Generalized Additive Models. *Statistical Science*, **1**, 297–310.

Hastie T., Tibshirani R., & Friedman J. (2009) *The Elements of Statistical Learning*. Springer, Berlin.

Hattab T., Ben Rais Lasram F., Albouy C., Sammari C., Romdhane M.S., Cury P., Leprieur F., & Le Loc'h F. (2013) The Use of a Predictive Habitat Model and a Fuzzy Logic Approach for Marine Management and Planning. *PLoS ONE*, **8**, e76430.

Hay C.H. (1990) The dispersal of sporophytes of *Undaria pinnatifida* by coastal shipping in New Zealand, and implications for further dispersal of Undaria in France. *British Phycological Journal*, **25**, 301–313.

Hayes M.A., Cryan P.M., & Wunder M.B. (2015) Seasonally-dynamic presence-only species distribution models for a cryptic migratory bat impacted by wind energy development. *PLoS ONE*, **10**, 1–20.

Heikkinen R.K., Marmion M., & Luoto M. (2012) Does the interpolation accuracy of species distribution models come at the expense of transferability? *Ecography*, **35**, 276–288.

Helmuth B., Mieszkowska N., Moore P., & Hawkins S.J. (2006) Living on the Edge of Two Changing Worlds: Forecasting the Responses of Rocky Intertidal Ecosystems to Climate Change. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 373–404.

Helmuth B., Yamane L., Lalwani S., Matzelle A., Tockstein A., & Gao N. (2011) Hidden signals of climate change in intertidal ecosystems: What (not) to expect when you are expecting. *Journal of Experimental Marine Biology and Ecology*, **400**, 191–199.

Herbrich R., Minka T., & Graepel T. (2006) TrueSkill: A Bayesian Skill Rating System. *Advances in Neural Information Processing Systems*, **20**, 569–576.

Higgs N.D. & Attrill M.J. (2015) Biases in biodiversity: wide-ranging species are discovered first in the deep sea. *Frontiers in Marine Science*, **2**.

Hijmans R.., Condori B., Carrillo R., & Kropff M.. (2003) A quantitative and constraint-specific method to assess the potential impact of new agricultural technology: the case of frost resistant potato for the Altiplano (Peru and Bolivia). *Agricultural Systems*, **76**, 895–911.

Hijmans R.J., Cameron S.E., Parra J.L., Jones P.G., & Jarvis A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hijmans R.J. & Graham C.H. (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, **12**, 2272–2281.

Hijmans R.J. (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, **93**, 679–688.

Hijmans R.J. (2016) *raster: Geographic Data Analysis and Modeling*. R package version 2.5-8. Available at: http://cran.r-project.org/package=raster.

Hijmans R.J., Phillips S., Leathwick J., & Elith J. (2016) *dismo: Species Distribution Modeling*. R package version 1.1-1. Available at: http://cran.r-project.org/package=dismo.

Hill A.W., Otegui J., Ariño A.H., & Guralnick R.P. (2010) *GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network.*

Hirzel A.H., Helfer V., & Metral F. (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, **145**, 111–121.

Hirzel A.H., Hausser J., Chessel D., & Perrin N. (2002) Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology*, **83**, 2027–2036.

Hirzel A.H., Le Lay G., Helfer V., Randin C., & Guisan A. (2006) Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, **199**, 142–152.

Holmes T.P., Aukema J.E., Von Holle B., Liebhold A., & Sills E. (2009) Economic Impacts of Invasive Species in Forests. *Annals of the New York Academy of Sciences*, **1162**, 18–38.

Hooper D.U., Chapin F.S., Ewel J.J. *et al.* (2005) Effects of biodiversity on ecosystem functioning: A consensus of current knowledge. *Ecological Monographs*, **75**, 3–35.

Hopfield J.J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, **79**, 2554–2558.

Howeth J.G., Gantz C.A., Angermeier P.L., Frimpong E.A., Hoff M.H., Keller R.P., Mandrak N.E., Marchetti M.P., Olden J.D., Romagosa C.M., & Lodge D.M. (2016) Predicting invasiveness of species in trade: climate match, trophic guild and fecundity influence establishment and impact of non-native freshwater fishes. *Diversity and Distributions*, **22**, 148–160.

Huang Z., Brooke B., & Li J. (2011) Performance of predictive models in marine benthic environments based on predictions of sponge distribution on the Australian continental shelf. *Ecological Informatics*, **6**, 205–216.

Huang M., Kong X., Varela S., & Duan R. (2016) The Niche Limitation Method (NicheLim), a new algorithm for generating virtual species to study biogeography. *Ecological Modelling*, **320**, 197–202.

Hui F., Warton D., Foster S., & Dunstan P. (2013) To mix or not to mix: comparing the predictive performance of mixture models versus separate species distribution models. *Ecology*, **94**, 1913–1919.

von Humboldt A. & Bonpland A. (1805) *Essai sur la géographie des plantes :accompagné d'un tableau physique des régions équinoxiales, fondé sur des mesures exécutées, depuis le*

dixième degré de latitude boréale jusqu'au dixième degré de latitude australe, pendant les années 1799,. Chez Levrault, Schoell et compagnie, libraires, Paris, FR.

Hurd C.L., Harrison P.J., Bischof K., & Lobban C.S. (2014) *Seaweed ecology and physiology.* Cambridge University Press, Cambridge, USA.

Hutchinson G.E. (1957) Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, **22**, 415–427.

ICES (2010) *ICES Biological Community dataset (DOME - Community). The International Council for the Exploration of the Sea, Copenhagen*. Available at: http://ecosystemdata.ices.dk/.

INVASIVES (2016) Available at: http://invasives.b.uib.no/.

Irigoyen A.J., Eyras C., & Parma A.M. (2011) Alien algae *Undaria pinnatifida* causes habitat loss for rocky reef fishes in north Patagonia. *Biological Invasions*, **13**, 17–24.

Janekovi F. & Novak T. (2012) PCA – A Powerful Method for Analyze Ecological Niches. *Principal Component Analysis - Multidisciplinary Applications* (ed. by P. Sanguansat), InTech, Rijeka, HRV.

Jiménez-Valverde A., Peterson A.T., Soberón J., Overton J.M., Aragón P., & Lobo J.M. (2011) Use of niche models in invasive species risk assessments. *Biological Invasions*, **13**, 2785–2797.

Johnston M.W. & Purkis S.J. (2014) Are lionfish set for a Mediterranean invasion? Modelling explains why this is unlikely to occur. *Marine Pollution Bulletin*, **88**, 138–147.

Jones C.G., Lawton J.H., & Shachak M. (1994) Organisms as Ecosystem Engineers. *Oikos*, **69**, 373–386.

de Jong Y., Verbeek M., Michelsen V. *et al.* (2014) Fauna Europaea – all European animal species on the web. *Biodiversity Data Journal*, **2**, e4034.

Jousson O., Pawlowski J., Zaninetti L., Meinesz A., & Boudouresque C. (1998) Molecular evidence for the aquarium origin of the green alga *Caulerpa taxifolia* introduced to the Mediterranean Sea. *Marine Ecology Progress Series*, **172**, 275–280.

Jueterbock A., Tyberghein L., Verbruggen H., Coyer J. a., Olsen J.L., & Hoarau G. (2013) Climate change impact on seaweed meadow distribution in the North Atlantic rocky intertidal. *Ecology and Evolution*, **3**, 1356–1373.

Kaschner K., Watson R., Trites A., & Pauly D. (2006) Mapping world-wide distributions of marine mammal species using a relative environmental suitability (RES) model. *Marine Ecology Progress Series*, **316**, 285–310.

Katsanevakis S., Zenetos A., Belchior C., & Cardoso A.C. (2013) Invading European Seas: Assessing pathways of introduction of marine aliens. *Ocean and Coastal Management*, **76**, 64–74.

Kearney M. & Porter W. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology letters*, **12**, 334–50.

Kearney M., Simpson S.J., Raubenheimer D., & Helmuth B. (2010) Modelling the ecological niche from functional traits. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 3469–3483.

Kramer-Schadt S., Niedballa J., Pilgrim J.D. *et al.* (2013) The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, **19**, 1366–1379.

Larson E.R., Gallagher R. V., Beaumont L.J., & Olden J.D. (2014) Generalized "avatar" niche shifts improve distribution models for invasive species. *Diversity and Distributions*, **20**, 1296–1306.

Lauzeral C., Grenouillet G., & Brosse S. (2012) Dealing with Noisy Absences to Optimize Species Distribution Models: An Iterative Ensemble Modelling Approach. *PLoS ONE*, **7**, e49508.

Leliaert F., Verbruggen H., Vanormelingen P., Steen F., López-Bautista J.M., Zuccarello G.C., & De Clerck O. (2014) DNA-based species delimitation in algae. *European Journal of Phycology*, **49**, 179–196.

Leroy B., Meynard C.N., Bellard C., & Courchamp F. (2016) virtualspecies, an R package to generate virtual species distributions. *Ecography*, **39**, 599–607.

Levin P.S., Coyer J.A., Petrik R., & Good T.P. (2002) Community-wide effects of nonindigenous species on temperate rocky reefs. *Ecology*, **83**, 3182–3193.

Leys C., Ley C., Klein O., Bernard P., & Licata L. (2013) Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, **49**, 764–766.

Li W. & Guo Q. (2013) How to assess the prediction accuracy of species presence-absence models without absence data? *Ecography*, **36**, 788-799.

Liaw A. & Wiener M. (2002) Classification and regression by randomForest. *R news*, **2**, 18–22.

Liu C., White M., & Newell G. (2011) Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography*, **34**, 232–243.

Liu C., White M., & Newell G. (2013) Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, **40**, 778–789.

Lobo J.M., Jiménez-Valverde A., & Real R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Lobo J.M., Jiménez-Valverde A., & Hortal J. (2010) The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, **33**, 103–114.

Lobo J.M. & Tognelli M.F. (2011) Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *Journal for Nature Conservation*, **19**, 1–7.

Lorena A.C., Jacintho L.F.O., Siqueira M.F., Giovanni R. De, Lohmann L.G., de Carvalho A.C.P.L.F., & Yamamoto M. (2011) Comparing machine learning classifiers in potential distribution modelling. *Expert Systems with Applications*, **38**, 5268–5275.

Lovell S.J., Stone S.F., & Fernandez L. (2006) The Economic Impacts of Aquatic Invasive Species: A Review of the Literature. *Agricultural and Resource Economics Review*, **35**, 195–208.

Lucy F., Roy H., Simpson A. *et al.* (2016) INVASIVESNET towards an International Association for Open Knowledge on Invasive Alien Species. *Management of Biological Invasions*, **7**, 131–139.

Lunetta R.S. & Lyon J.G. (2004) *Remote Sensing and GIS Accuracy Assessment.* CRC Press.

Lüning K. (1990) *Seaweeds: Their Environment, Biogeography, and Ecophysiology.* John Wiley & Sons, New York, New York, USA.

Ma K.C.K., Deibel D., Law K.K.M., Aoki M., McKenzie C.H., & Palomares M.L.D. (2017) Richness and zoogeography of ascidians (Tunicata: Ascidiacea) in eastern Canada. *Canadian Journal of Zoology*, **95**, 51–59.

MacLeod C.D., Mandleberg L., Schweder C., Bannon S.M., & Pierce G.J. (2008) A comparison of approaches for modelling the occurrence of marine animals. *Hydrobiologia*, **612**, 21–32.

Madon B., Warton D.I., & Araújo M.B. (2013) Community-level vs species-specific approaches to model selection. *Ecography*, **36**, 1291-1298.

Magner L.N. (2002) *A History of the Life Sciences, Revised and Expanded.* CRC Press, Abingdon, UK.

Marcelino V.R. & Verbruggen H. (2015) Ecological niche models of invasive seaweeds. *Journal of Phycology*, **51**, 606–620.

Márcia Barbosa A., Real R., Muñoz A.-R., & Brown J. a. (2013) New measures for assessing model equilibrium and prediction mismatch in species distribution models. *Diversity and Distributions*, **19**, 1333-1338.

Marmion M., Parviainen M., Luoto M., Heikkinen R.K., & Thuiller W. (2009) Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, **15**, 59–69.

Martínez B., Arenas F., Trilla A., Viejo R.M., & Carreño F. (2015) Combining physiological threshold knowledge to species distribution models is key to improving forecasts of the future niche for macroalgae. *Global Change Biology*, **21**, 1422–1433.

Martins I., Oliveira J.M., Flindt M.R., & Marques J.C. (1999) The effect of salinity on the growth rate of the macroalgae *Enteromorpha intestinalis* (Chlorophyta) in the Mondego estuary (west Portugal). *Acta Oecologica*, **20**, 259–265.

Mayr E. (1985) *The growth of biological thought: diversity, evolution and inheritance.* The Belknap Press of Harvard Univ. Press, Cambridge, USA.

Mazza G., Aquiloni L., Inghilesi A., Giuliani C., Lazzaro L., Ferretti G., Lastrucci L., Foggi B., & Tricarico E. (2015) Aliens just a click away: the online aquarium trade in Italy. *Management of Biological Invasions*, **6**, 253–261.

McCulloch W.S. & Pitts W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, 115–133.

McDevit D.C. & Saunders G.W. (2009) On the utility of DNA barcoding for species differentiation among brown macroalgae (Phaeophyceae) including a novel extraction protocol. *Phycological Research*, **57**, 131–141.

McDonald J.I., Huisman J.M., Hart F.N., Dixon R.R.M., Lewis J.A., & others (2015) The first detection of the invasive macroalga *Codium fragile* subsp. *fragile* (Suringar) Hariot in Western Australia. *BioInvasions Records*, **4**, 75–80.

McGeoch M.A., Butchart S.H.M., Spear D., Marais E., Kleynhans E.J., Symes A., Chanson J., & Hoffmann M. (2010) Global indicators of biological invasion: Species numbers, biodiversity impact and policy responses. *Diversity and Distributions*, **16**, 95–108.

McIvor L., Maggs C.A., Provan J., & Stanhope M.J. (2001) rbcL sequences reveal multiple cryptic introductions of the Japanese red alga *Polysiphonia harveyi*. *Molecular Ecology*, **10**, 911–919.

Meinesz A. (1999) *Killer algae: a true tale of biological invasion.* Uinversity of Chicago Press, Chicago.

Mendoza R., Luna S., & Aguilera C. (2015) Risk assessment of the ornamental fish trade in Mexico: analysis of freshwater species and effectiveness of the FISK (Fish Invasiveness Screening Kit). *Biological Invasions*, **17**, 3491–3502.

Merckx B., Steyaert M., Vanreusel A., Vincx M., & Vanaverbeke J. (2011) Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. *Ecological Modelling*, **222**, 588–597.

Merow C., Smith M.J., Edwards T.C., Guisan A., McMahon S.M., Normand S., Thuiller W., Wüest R.O., Zimmermann N.E., & Elith J. (2014) What do we gain from simplicity versus complexity in species distribution models? *Ecography*, **37**, 1267–1281.

Merow C., Smith M.J., & Silander J.A. (2013) A practical guide to MaxEnt for modeling species' distributions : what it does , and why inputs and settings matter. *Ecography*, **36**, 1058–1069.

Meynard C.N. & Kaplan D.M. (2013) Using virtual species to study species distributions and model performance. *Journal of Biogeography*, **40**, 1–8.

Meynard C.N. & Quinn J.F. (2007) Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, **34**, 1455–1469.

Mieszkowska N., Milligan G., Burrows M.T., Freckleton R., & Spencer M. (2013) Dynamic species distribution models from categorical survey data. *Journal of Animal Ecology*, **82**, 1215–1226.

Miller J. (1991) Short report: Reaction time analysis with outlier exclusion: Bias varies with sample size. *The Quarterly Journal of Experimental Psychology Section A*, **43**, 907–912.

Miller J.A. (2014) Virtual species distribution models: Using simulated data to evaluate aspects of model performance. *Progress in Physical Geography*, **38**, 117–128.

Mineur F., Belsher T., Johnson M.P., Maggs C.A., & Verlaque M. (2007a) Experimental assessment of oyster transfers as a vector for macroalgal introductions. *Biological Conservation*, **137**, 237–247.

Mineur F., Johnson M.P., Maggs C.A., & Stegenga H. (2007b) Hull fouling on commercial ships as a vector of macroalgal introduction. *Marine Biology*, **151**, 1299–1307.

Mineur F., Johnson M.P., & Maggs C.A. (2008) Macroalgal introductions by hull fouling on recreational vessels: Seaweeds and sailors. *Environmental Management*, **42**, 667–676.

Mineur F., Davies A.J., Maggs C. a, Verlaque M., & Johnson M.P. (2010) Fronts, jumps and secondary introductions suggested as different invasion patterns in marine species, with an increase in spread rates over time. *Proceedings. Biological sciences / The Royal Society*, **277**, 2693–701.

Mineur F., Cook E.J., Minchin D., Bohn K., MacLeod A., & Maggs C.A. (2012) Changing coasts: marine aliens and artificial structures. *Oceanography and Marine Biology: An Annual Review* (ed. by R.N. Gibson, R.J.A. Atkinson, J.D.M. Gordon, and R.N. Hughes), pp. 189–234. CRC Press, Abingdon, UK.

Mineur F., Le Roux A., Maggs C.A., & Verlaque M. (2014) Positive Feedback Loop between Introductions of Non-Native Marine Species and Cultivation of Oysters in Europe. *Conservation Biology*, **28**, 1667–1676.

Mineur F., Arenas F., Assis J. *et al.* (2015) European seaweeds under pressure: Consequences for communities and ecosystem functioning. *Journal of Sea Research*, **98**, 91–108.

Mitchell C.E., Agrawal A.A., Bever J.D. *et al.* (2006) Biotic interactions and plant invasions. *Ecology Letters*, **9**, 726–740.

Molnar J.L., Gamboa R.L., Revenga C., & Spalding M.D. (2008) Assessing the global threat of invasive species to marine biodiversity. *Frontiers in Ecology and the Environment*, **6**, 485–492.

Monk J. (2013) How long should we ignore imperfect detection of species in the marine environment when modelling their distribution? *Fish and Fisheries*, **15**, 352-358.

Monteiro C.A., Engelen A.H., & Santos R.O.P. (2009) Macro- and mesoherbivores prefer native seaweeds over the invasive brown seaweed *Sargassum muticum*: A potential regulating role on invasions. *Marine Biology*, **156**, 2505–2515.

Moreno-Amat E., Mateo R.G., Nieto-Lugilde D., Morueta-Holme N., Svenning J.C., & García-Amorena I. (2015) Impact of model complexity on cross-temporal transferability in Maxent species distribution models: An assessment using paleobotanical data. *Ecological Modelling*, **312**, 308–317.

Mouton A.M., De Baets B., Van Broekhoven E., & Goethals P.L.M. (2009) Prevalence-adjusted optimisation of fuzzy models for species distribution. *Ecological Modelling*, **220**, 1776–1786.

Mouton A.M., De Baets B., & Goethals P.L.M. (2010) Ecological relevance of performance criteria for species distribution models. *Ecological Modelling*, **221**, 1995–2002.

Naimi B., Skidmore A.K., Groen T. a., & Hamm N. a. S. (2011) Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography*, **38**, 1497–1509.

Naimi B., Hamm N. a. S., Groen T. a., Skidmore A.K., & Toxopeus A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, **37**, 191–203.

Naimi B. & Araújo M.B. (2016) sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography*, **39**, 368–375.

Negahban S., Oh S., & Shah D. (2017) Rank Centrality: Ranking from Pairwise Comparisons. *Operations Research*, **65**, 266–287.

Neill P.E., Alcalde O., Faugeron S., Navarrete S.A., & Correa J.A. (2006) Invasion of *Codium fragile* ssp. *tomentosoides* in northern Chile: A new threat for Gracilaria farming. *Aquaculture*, **259**, 202–210.

Nejrup L.B., Staehr P.A., & Thomsen M.S. (2013) Temperature- and light-dependent growth and metabolism of the invasive red algae *Gracilaria vermiculophylla* – a comparison with two native macroalgae. *European Journal of Phycology*, **48**, 295–308.

Nelder J.A. & Wedderburn R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, 370–384.

Nenzén H.K., Swab R.M., Keith D. a., & Araújo M.B. (2012) demoniche - an R-package for simulating spatially-explicit population dynamics. *Ecography*, **35**, 577–580.

Nikolić V., Žuljević A., Antolić B., Despalatović M., & Cvitković I. (2010) Distribution of invasive red alga *Womersleyella setacea* (Hollenberg) RE Norris (Rhodophyta, Ceramiales) in the Adriatic Sea. *Acta Adriatica*, **51**, 195–202.

Norse E.A. (1993) *Global marine biological diversity: a strategy for building conservation into decision making.* Island Press, Washington, USA.

Nyberg C.D. & Wallentinus I. (2005) Can species traits be used to predict marine macroalgal introductions? *Biological Invasions*, **7**, 265–279.

Nyström Sandman A., Wikström S. a., Blomqvist M., Kautsky H., & Isaeus M. (2013) Scale-dependent influence of environmental variables on species distribution: a case study on five coastal benthic species in the Baltic Sea. *Ecography*, **36**, 354–363.

O'Connor R.J.R.J. (2002) The conceptual basis of species distribution modelling; time for paradigm shift? *Predicting Species Occurrences: Issues of Accuracy and Scale* pp. 25–33. Island Press, Covelo, CA.

O'Dor R., Miloslavich P., & Yarincik K. (2010) Marine Biodiversity and Biogeography – Regional Comparisons of Global Issues, an Introduction. *PLoS ONE*, **5**, e11871.

O'Hara R.B. & Sillanpää M.J. (2009) A review of bayesian variable selection methods: What, how and which. *Bayesian Analysis*, **4**, 85–118.

Obura D. (2012) The Diversity and Biogeography of Western Indian Ocean Reef-Building Corals. *PLoS ONE*, **7**, e45013.

Odom R.L. & Walters L.J. (2014) A safe alternative to invasive Caulerpa taxifolia (Chlorophtya)? Assessing aquarium-release invasion potential of aquarium strains of the macroalgal genus Chaetomorpha (Chlorophyta). *Biological Invasions*, **16**, 1589–1597.

Oksanen J., Blanchet F.G., Friendly M. *et al.* (2017) *vegan: Community Ecology Package*. R package version 2.4-3. Available at: https://cran.r-project.org/package=vegan.

Olden J.D., Joy M.K., & Death R.G. (2004) An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, **178**, 389–397.

Otto T., Vasconcellos E., Gomes L., Moreira A., Degrave W., Mendonça-Lima L., & Alves-Ferreira M. (2008) ChromaPipe: a pipeline for analysis, quality control and management for a DNA sequencing facility. *Genet Mol Res.*, **7**, 861–871.

Ovaskainen O., Abrego N., Halme P., & Dunson D. (2015) Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*, **7**, 549-555.

Pacifici K., Reich B.J., Miller D.A.W., Gardner B., Stauffer G., Singh S., McKerrow A., & Collazo J.A. (2017) Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology*, **98**, 840–850.

Padilla D.K. & Williams S.L. (2004) Beyond ballast water: aquarium and ornamental trades as sources of invasive species in aquatic ecosystems. *Frontiers in Ecology and the Environment*, **2**, 131–138.

Palialexis A., Georgakarakos S., Karakassis I., Lika K., & Valavanis V.D. (2011) Prediction of marine species distribution from presence–absence acoustic data: comparing the fitting efficiency and the predictive capacity of conventional and novel distribution models. *Hydrobiologia*, **670**, 241–266.

Palomares M.L.D., Pauly D., & Editors (2017) SeaLifeBase. Available from http://sealifebase.org. Accessed 2017-05-11.

Parmesan C. & Yohe G. (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, **421**, 37–42.

Pauly K., Jupp B.P., & de Clerck O. (2011) Modelling the distribution and ecology of Trichosolen blooms on coral reefs worldwide. *Marine Biology*, **158**, 2239–2246.

Pearce J.L., Cherry K., M. D., S. F., & Whish G. (2001) Incorporating expert opinion and fine-scale vegetation mapping into statistical models of faunal distribution. *Journal of Applied Ecology*, **38**, 412–424.

Pearce J.L. & Boyce M.S. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, **43**, 405–412.

Pearman P.B., Guisan A., Broennimann O., & Randin C.F. (2008) Niche dynamics in space and time. *Trends in Ecology and Evolution*, **23**, 149–158.

Pearson R.G. & Dawson T.P. (2003) Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? *Global Ecology and Biogeography*, **12**, 361–371.

Pearson R.G., Phillips S.J., Loranty M.M., Beck P.S. a., Damoulas T., Knight S.J., & Goetz S.J. (2013) Shifts in Arctic vegetation and associated feedbacks under climate change. *Nature Climate Change*, **3**, 673–677.

Pereira H.M., Leadley P.W., Proenca V. *et al.* (2010) Scenarios for Global Biodiversity in the 21st Century. *Science*, **330**, 1496–1501.

Perry A.L. (2005) Climate Change and Distribution Shifts in Marine Fishes. *Science*, **308**, 1912–1915.

Peterson A.T., Soberón J., Pearson R.G., Anderson R.P., Martínez-Meyer E., Nakamura M., & Araújo M.B. (2011) *Ecological niches and geographic distributions (MPB-49).* Princeton University Press, Princeton, USA.

Petitpierre B., Kueffer C., Broennimann O., Randin C., Daehler C., & Guisan A. (2012) Climatic Niche Shifts Are Rare Among Terrestrial Plant Invaders. *Science*, **335**, 1344–1348.

Petitpierre B., Broennimann O., Kueffer C., Daehler C., & Guisan A. (2017) Selecting predictors to maximize the transferability of species distribution models: lessons from cross-continental plant invasions. *Global Ecology and Biogeography*, **26**, 275–287.

Phillips S.J., Dudík M., & Schapire R.E. (2004) A maximum entropy approach to species distribution modeling. *Twenty-first international conference on Machine learning - ICML '04*, 655–662.

Phillips S.J. & Dudík M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

Phillips S.J., Dudík M., Elith J., Graham C.H., Lehmann A., Leathwick J., & Ferrier S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–197.

Pittman S.J. & Brown K. a (2011) Multi-scale approach for predicting fish species distributions across coral reef seascapes. *PloS one*, **6**, e20583.

Pollock L.J., Tingley R., Morris W.K., Golding N., O'Hara R.B., Parris K.M., Vesk P.A., & Mccarthy M.A. (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, **5**, 397–406.

De Pooter D., Appeltans W., Bailly N. *et al.* (2017) Toward a new data standard for combined marine biological and environmental datasets - expanding OBIS beyond species occurrences. *Biodiversity Data Journal*, **5**, e10989.

Pratheepa M., Verghese A., & Bheemanna H. (2016) Shannon information theory a useful tool for detecting significant abiotic factors influencing the population dynamics of Helicoverpa armigera (Hübner) on cotton crop. *Ecological Modelling*, **337**, 25–28.

Provan J., Murphy S., & Maggs C.A. (2004) Tracking the invasive history of the green alga *Codium fragile* ssp. *tomentosoides*. *Molecular Ecology*, **14**, 189–194.

Provan J., Booth D., Todd N.P., Beatty G.E., & Maggs C.A. (2008) Tracking biological invasions in space and time: elucidating the invasive history of the green alga *Codium fragile* using old DNA. *Diversity and Distributions*, **14**, 343–354.

Provoost P., Bosch S., & Appeltans W. (2016) *robis: R client for the OBIS API*. R package version 0.1.5. Available at: https://github.com/iobis/robis.

Pruesse E., Quast C., Knittel K., Fuchs B.M., Ludwig W., Peplies J., & Glöckner F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research*, **35**, 7188–96.

Pulliam H.R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**, 349–361.

Purves D., Scharlemann J., Harfoot M., Newbold T., Tittensor D.P., Hutton J., & Emmott S. (2013) Ecosystems: Time to model all life on Earth. *Nature*, **493**, 295–7.

Qiao H., Peterson A.T., Campbell L.P., Soberón J., Ji L., & Escobar L.E. (2016) NicheA: creating virtual species and ecological niches in multivariate environmental scenarios. *Ecography*, **39**, 805–813.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Available at: http://www.r-project.org/.

Radosavljevic A. & Anderson R.P. (2014) Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography*, **41**, 629–643.

Raes N. & ter Steege H. (2007) A null-model for significance testing of presence-only species distribution models. *Ecography*, **30**, 727–736.

Ranc N., Santini L., Rondinini C., Boitani L., Poitevin F., Angerbjörn A., & Maiorano L. (2016) Performance tradeoffs in target-group bias correction for species distribution models. *Ecography*, Early View.

Randin C.F., Dirnböck T., Dullinger S., Zimmermann N.E., Zappa M., & Guisan A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography*, **33**, 1689–1703.

Rapacciuolo G., Roy D.B., Gillings S., Fox R., Walker K., & Purvis A. (2012) Climatic Associations of British Species Distributions Show Good Transferability in Time but Low Predictive Accuracy for Range Change. *PLoS ONE*, **7**, e40212.

Ray G.C. (1996) Biodiversity is biogeography: implications for conservation. *Oceanography*, **9**, 50–59.

Ready J., Kaschner K., South A.B., Eastwood P.D., Rees T., Rius J., Agbayani E., Kullander S., & Froese R. (2010) Predicting the distributions of marine organisms at the global scale. *Ecological Modelling*, **221**, 467–478.

Real R., Barbosa A.M., & Vargas J.M. (2006) Obtaining Environmental Favourability Functions from Logistic Regression. *Environmental and Ecological Statistics*, **13**, 237–245.

Reineking B. & Schröder B. (2006) Constrain to perform: Regularization of habitat models. *Ecological Modelling*, **193**, 675–690.

Reiss H., Birchenough S., Borja A., Buhl-Mortensen L., Craeymeersch J., Dannheim J., Darr A., Galparsoro I., Gogina M., Neumann H., Populus J., Rengstorf A.M., Valle M., van Hoey G., Zettler M.L., & Degraer S. (2015) Benthos distribution modelling and its relevance for marine ecosystem management. *ICES Journal of Marine Science*, **72**, 297–315.

Richardson D.M. & Whittaker R.J. (2010) Conservation biogeography - foundations, concepts and challenges. *Diversity and Distributions*, **16**, 313–320.

Rixon C.A.M., Duggan I.C., Bergeron N.M.N., Ricciardi A., & Macisaac H.J. (2005) Invasion risks posed by the aquarium trade and live fish markets on the Laurentian Great Lakes. *Biodiversity and Conservation*, **14**, 1365–1381.

Roberts D.R., Bahn V., Ciuti S. *et al.* (2016) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*.

Robertson D.R. (2008) Global biogeographical data bases on marine fishes: caveat emptor. *Diversity and Distributions*, **14**, 891–892.

Robertson T., Döring M., Guralnick R., Bloom D., Wieczorek J., Braak K., Otegui J., Russell L., & Desmet P. (2014) The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. *PLoS ONE*, **9**, e102623.

Robertson M.P., Visser V., & Hui C. (2016) Biogeo: an R package for assessing and improving data quality of occurrence record datasets. *Ecography*, **39**, 394–401.

Robinson L.M., Elith J., Hobday a. J., Pearson R.G., Kendall B.E., Possingham H.P., & Richardson a. J. (2011) Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities. *Global Ecology and Biogeography*, **20**, 789–802.

Rocchini D., Hortal J., Lengyel S., Lobo J.M., Jimenez-Valverde A., Ricotta C., Bacaro G., & Chiarucci A. (2011) Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography*, **35**, 211–226.

Rödder D. & Lötters S. (2009) Niche shift versus niche conservatism? Climatic characteristics of the native and invasive ranges of the Mediterranean house gecko (*Hemidactylus turcicus*). *Global Ecology and Biogeography*, **18**, 674–687.

Royle J.A., Chandler R.B., Yackulic C., & Nichols J.D. (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, **3**, 545–554.

Samperio-Ramos G., Olsen Y.S., Tomas F., & Marbà N. (2015) Ecophysiological responses of three Mediterranean invasive seaweeds (*Acrothamnion preissii*, *Lophocladia lallemandii* and *Caulerpa cylindracea*) to experimental warming. *Marine Pollution Bulletin*, **96**, 418–423.

Santika T. & Hutchinson M.F. (2009) The effect of species response form on species distribution model prediction and inference. *Ecological Modelling*, **220**, 2365–2379.

Saunders G.W. (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1879–1888.

Saunders G. & Kucera H. (2010) An evaluation of rbcL, tufA, UPA, LSU and ITS as DNA barcode markers for the marine green macroalgae. *Cryptogam Algol*, **31**, 487–528.

Saunders G.W. & Moore T.E. (2013) Refinements for the amplification and sequencing of red algal DNA barcode and RedToL phylogenetic markers: a summary of current primers, profiles and strategies. *ALGAE*, **28**, 31–43.

Saupe E.E., Barve V., Myers C.E., Soberón J., Barve N., Hensz C.M., Peterson A.T., Owens H.L., & Lira-Noriega A. (2012) Variation in niche and distribution model performance: The need for a priori assessment of key causal factors. *Ecological Modelling*, **237–238**, 11–22.

Sbrocco E.J. & Barber P.H. (2013) MARSPEC: ocean climate layers for marine spatial ecology. *Ecology*, **94**, 979.

Schaffelke B., Smith J.E., & Hewitt C.L. (2006) Introduced macroalgae - A growing concern. *Journal of Applied Phycology*, **18**, 529–541.

Schaffelke B. & Hewitt C.L. (2007) Impacts of introduced seaweeds. *Botanica Marina*, **50**, 397–417.

Scheibling R. & Gagnon P. (2006) Competitive interactions between the invasive green alga *Codium fragile* ssp. *tomentosoides* and native canopy-forming seaweeds in Nova Scotia (Canada). *Marine Ecology Progress Series*, **325**, 1–14.

Seebens H., Blackburn T.M., Dyer E.E. *et al*. (2017) No saturation in the accumulation of alien species worldwide. *Nature Communications*, **8**, 14435.

Seebens H., Gastner M.T., & Blasius B. (2013) The risk of marine bioinvasion caused by global shipping. *Ecology Letters*, **16**, 782–790.

Selama O., James P., Nateche F., Wellington E.M.H., & Hacène H. (2013) The World Bacterial Biogeography and Biodiversity through Databases: A Case Study of NCBI Nucleotide Database and GBIF Database. *BioMed Research International*, **2013**, 1–11.

Senay S.D., Worner S.P., & Ikeda T. (2013) Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. *PLoS ONE*, **8**, e71218.

Seoane J., Bustamante J., & Diaz-delgado R. (2005) Effect of Expert Opinion on the Predictive Ability of Environmental Models of Bird Distribution. *Conservation Biology*, **19**, 512–522.

Shcheglovitova M. & Anderson R.P. (2013) Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecological Modelling*, **269**, 9–17.

Sherwood M.K. (1991) Quality assurance in biomedical or clinical engineering. *Journal of clinical engineering*, **16**, 479–83.

Šiaulys A. & Bučas M. (2012) Species distribution modelling of benthic invertebrates in the south-eastern Baltic Sea. *Baltica*, **25**, 163–170.

Smith A.B. (2013) On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. *Diversity and Distributions*, **19**, 867–872.

Smith A.B., Santos M.J., Koo M.S., Rowe K.M.C., Rowe K.C., Patton J.L., Perrine J.D., Beissinger S.R., & Moritz C. (2013) Evaluation of species distribution models by resampling of sites surveyed a century ago by Joseph Grinnell. *Ecography*, **36**, 1017–1031.

Soberón J. (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, **10**, 1115–1123.

Soberón J. & Peterson A.T. (2005) Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas. *Biodiversity Informatics*, **2**, 1–10.

Sorte C.J.B., Williams S.L., & Zerebecki R.A. (2010) Ocean warming increases threat of invasive species in a marine fouling community. *Ecology*, **91**, 2198–2204.

Sousa R., Gutiérrez J.L., & Aldridge D.C. (2009) Non-indigenous invasive bivalves as ecosystem engineers. *Biological Invasions*, **11**, 2367–2385.

Spalding M.D., Fox H.E., Allen G.R. *et al.* (2007) Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas. *BioScience*, **57**, 573.

Stam W.T., Olsen J.L., Zaleski S.F., Murray S.N., Brown K.R., & Walters L.J. (2006) A forensic and phylogenetic survey of *Caulerpa* species (Caulerpales, Chlorophyta) from the Florida coast, local aquarium shops, and e-commerce: establishing a proactive baseline for early detection. *Journal of Phycology*, **42**, 1113–1124.

Steen F., Aragay J., Zuljevic A., Verbruggen H., Mancuso F.P., Bunker F., Vitales D., Gómez Garreta A., & De Clerck O. (2017) Tracing the introduction history of the brown seaweed *Dictyota cyanoloma* (Phaeophyceae, Dictyotales) in Europe. *European Journal of Phycology*, **52**, 31–42.

Stegenga H. & Karremans M. (2015) Overzicht van de roodwier-exoten in de mariene wateren van Zuidwest-Nederland. *Gorteria*, **37**, 141–157.

Stirling D.A., Boulcott P., Scott B.E., & Wright P.J. (2016) Using verified species distribution models to inform the conservation of a rare marine species. *Diversity and Distributions*, **22**, 808–822.

Stockwell D.R.B. & Noble I.R. (1992) Induction of sets of rules from animal distribution data: A robust and informative method of data analysis. *Mathematics and Computers in Simulation*, **33**, 385–390.

Stokland J.N., Halvorsen R., & Støa B. (2011) Species distribution modelling—Effect of design and sample size of pseudo-absence observations. *Ecological Modelling*, **222**, 1800–1809.

Stoklosa J., Daly C., Foster S.D., Ashcroft M.B., & Warton D.I. (2015) Spatial models for distance sampling data: recent developments and future directions. *Methods in Ecology and Evolution*, **4**, 1001–1010.

Strain E.M.A., Thomson R.J., Micheli F., Mancuso F.P., & Airoldi L. (2014) Identifying the interacting roles of stressors in driving the global loss of canopy-forming to mat-forming algae in marine ecosystems. *Global Change Biology*, **20**, 3300–3312.

Strecker A.L., Campbell P.M., & Olden J.D. (2011) The Aquarium Trade as an Invasion Pathway in the Pacific Northwest. *Fisheries*, **36**, 74–85.

Streftaris N. & Zenetos A. (2006) Alien Marine Species in the Mediterranean - the 100 "Worst Invasives" and their Impact. *Mediterranean Marine Science*, **7**.

Strubbe D., Broennimann O., Chiron F., & Matthysen E. (2013) Niche conservatism in non-native birds in Europe: Niche unfilling rather than niche expansion. *Global Ecology and Biogeography*, **22**, 962–970.

Syfert M.M., Smith M.J., & Coomes D. a (2013) The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. *PLoS ONE*, **8**, e55158.

Synes N.W. & Osborne P.E. (2011) Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change. *Global Ecology and Biogeography*, **20**, 904–914.

Syphard A.D., Radeloff V.C., Keuler N.S., Taylor R.S., Hawbaker T.J., Stewart S.I., & Clayton M.K. (2008) Predicting spatial patterns of fire on a southern California landscape. *International Journal of Wildland Fire*, **17**, 602.

Tamura K., Stecher G., Peterson D., Filipski A., & Kumar S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*, **30**, 2725–2729.

Theobald E.J., Ettinger A.K., Burgess H.K., DeBey L.B., Schmidt N.R., Froehlich H.E., Wagner C., HilleRisLambers J., Tewksbury J., Harsch M.A., & Parrish J.K. (2015) Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation*, **181**, 236–244.

Thomsen M.S. & McGlathery K.J. (2007) Stress tolerance of the invasive macroalgae *Codium fragile* and *Gracilaria vermiculophylla* in a soft-bottom turbid lagoon. *Biological Invasions*, **9**, 499–513.

Thuiller W., Lavorel S., Araujo M.B., Sykes M.T., & Prentice I.C. (2005) Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences*, **102**, 8245–8250.

Thuiller W., Lafourcade B., Engler R., & Araújo M.B. (2009) BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

Thuiller W., Georges D., Engler R., & Breiner F. (2016) *biomod2: Ensemble Platform for Species Distribution Modeling*. R package version 3.3-7. Available at: http://cran.r-project.org/package=biomod2.

Tikhonov G., Abrego N., Dunson D., & Ovaskainen O. (2017) Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods in Ecology and Evolution*, **8**, 443–452.

Title P.O. & Bemmels J.B. (2017) envirem : An expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography*.

Tittensor D.P., Mora C., Jetz W., Lotze H.K., Ricard D., Berghe E. Vanden, & Worm B. (2010) Global patterns and predictors of marine biodiversity across taxa. *Nature*, **466**, 1098–101.

Torres L., Read A., & Halpin P. (2008) Fine-scale habitat modeling of a top marine predator: do prey data improve predictive capacity. *Ecological Applications*, **18**, 1702–1717.

Tronholm A., Steen F., Tyberghein L., Leliaert F., Verbruggen H., Antonia Ribera Siguan M., & De Clerck O. (2010) Species delimitation, taxonomy, and biogeography of *Dictyota* in Europe (Dictyotales, Phaeophyceae). *Journal of Phycology*, **46**, 1301–1321.

Tsoar A., Allouche O., Steinitz O., Rotem D., & Kadmon R. (2007) A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, **13**, 397–405.

Tumer K. & Ghosh J. (1996) Error Correlation and Error Reduction in Ensemble Classifiers. *Connection Science*, **8**, 385–404.

Tyberghein L., Verbruggen H., Pauly K., Troupin C., Mineur F., & De Clerck O. (2012) Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography*, **21**, 272–281.

Václavík T. & Meentemeyer R.K. (2009) Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling*, **220**, 3248–3258.

Vandepitte L., Hernandez F., Claus S., Vanhoorne B., Hauwere N., Deneudt K., Appeltans W., & Mees J. (2011) Analysing the content of the European Ocean Biogeographic Information System (EurOBIS): available data, limitations, prospects and a look at the future. *Hydrobiologia*, **667**, 1–14.

Vandepitte L., Vanhoorne B., Kraberg A. *et al.* (2010) Data integration for European marine biodiversity research: creating a database on benthos and plankton to study large-scale patterns and long-term changes. *Hydrobiologia*, **644**, 1–13.

VanDerWal J., Shoo L.P., Graham C., & Williams S.E. (2009) Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, **220**, 589–594.

Varela S., Anderson R.P., García-Valdés R., & Fernández-González F. (2014) Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, **37**, 1084–1091.

Veloz S.D. (2009) Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, **36**, 2290–2299.

Verbruggen H., Brookes M.J.L., & Costa J.F. (2016) DNA barcodes and morphometric data indicate that *Codium fragile* (Bryopsidales, Chlorophyta) may consist of two species. *Phycologia*, **56**, 54–62.

Verbruggen H., Leliaert F., Maggs C.A., Shimada S., Schils T., Provan J., Booth D., Murphy S., De Clerck O., Littler D.S., Littler M.M., & Coppejans E. (2007) Species boundaries and phylogenetic relationships within the green algal genus *Codium* (Bryopsidales) based on plastid DNA sequences. *Molecular Phylogenetics and Evolution*, **44**, 240–254.

Verbruggen H., Tyberghein L., Belton G.S., Mineur F., Jueterbock A., Hoarau G., Gurgel C.F.D., & De Clerck O. (2013) Improving Transferability of Introduced Species' Distribution Models: New Tools to Forecast the Spread of a Highly Invasive Seaweed. *PLoS ONE*, **8**, e68337.

Vergés A., Comalada N., Sánchez N., & Brodie J. (2013) A reassessment of the foliose Bangiales (Rhodophyta) in the Balearic Islands including the proposed synonymy of *Pyropia olivii* with *Pyropia koreana*. *Botanica Marina*, **56**.

Vergés A., Steinberg P.D., Hay M.E. *et al.* (2014) The tropicalization of temperate marine ecosystems: climate-mediated changes in herbivory and community phase shifts. *Proceedings of the Royal Society B: Biological Sciences*, **281**.

Verlaque M., Boudouresque C.F., Meinesz A., & Gravez V. (2000) The *Caulerpa racemosa* Complex (Caulerpales, Ulvophyceae) in the Mediterranean Sea. *Botanica Marina*, **43**.

Verlaque M., Boudouresque C.F., & Mineur F. (2007) Oyster transfers as a vector for marine species introductions: a realistic approach based on macrophytes. *CIESM Workshop Monographs, Monaco. Vol. 32. 2007.* pp. 39–48.

Verlaque M., Ruitton S., Mineur F., & Boudouresque C.F. (2015) *Macrophytes.* CIESM Publishers, Monaco, FR.

Vilà M., Basnou C., Pyšek P., Josefsson M., Genovesi P., Gollasch S., Nentwig W., Olenin S., Roques A., Roy D., & Hulme P.E. (2010) How well do we understand the impacts of alien species on ecosystem services? A pan-European, cross-taxa assessment. *Frontiers in Ecology and the Environment*, **8**, 135–144.

Wallace A.R. (1876) *The geographical distribution of animals, with a study of the relations of living and extinct faunas as elucidating the past changes of the earth's surface.* Macmillan and Co., London, UK.

Wallentinus I. (2002) Introduced marine algae and vascular plants in European aquatic environments. *Invasive aquatic species of Europe. Distribution, impacts and management* (ed. by E. Leppäkoski, S. Gollasch, and S. Olenin), pp. 27–52. Springer Netherlands, Amsterdam.

Walters L.J., Brown K.R., Stam W.T., & Olsen J.L. (2006) E-commerce and Caulerpa: unregulated dispersal of invasive species. *Frontiers in Ecology and the Environment*, **4**, 75–79.

Walther G.-R., Post E., Convey P., Menzel A., Parmesan C., Beebee T.J.C., Fromentin J., Hoegh-Guldberg O., & Bairlein F. (2002) Ecological responses to recent climate change. *Nature*, **416**, 389–395.

Walther G.-R., Roques A., Hulme P.E. *et al.* (2009) Alien species in a warmer world: risks and opportunities. *Trends in ecology & evolution*, **24**, 686–93.

Ward D.F. (2006) Modelling the potential geographic distribution of invasive ant species in New Zealand. *Biological Invasions*, **9**, 723–735.

Warren D. & Seifert S. (2011) Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications*, **21**, 335–342.

Warton D.I. & Shepherd L.C. (2010) Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, **4**, 1383–1402.

Wasof S., Lenoir J., Aarrestad P.A. *et al.* (2015) Disjunct populations of European vascular plant species keep the same climatic niches. *Global Ecology and Biogeography*, **24**, 1401–1412.

Watson H.C. (1847) *Cybele Britannica; or British plants and their geographical relations.* Longman & co., London, UK.

Webber B.L., Le Maitre D.C., & Kriticos D.J. (2012) Comment on "Climatic Niche Shifts Are Rare Among Terrestrial Plant Invaders." *Science*, **338**, 193–193.

Wenger S.J. & Olden J.D. (2012) Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, **3**, 260–267.

Wernberg T., Bennett S., Babcock R.C. *et al.* (2016) Climate-driven regime shift of a temperate marine ecosystem. *Science*, **353**, 169–172.

Whitfield P., Gardner T., Vives S., Gilligan M., Courtenay Ray W., Ray G., & Hare J. (2002) Biological invasion of the Indo-Pacific lionfish *Pterois volitans* along the Atlantic coast of North America. *Marine Ecology Progress Series*, **235**, 289–297.

Whittaker R.H., Levin S.A., & Root R.B. (1973) Niche, Habitat, and Ecotope. *The American naturalist*, **107**, 321–338.

Wickham H., Chang W., & RStudio (2016) *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 2.2.1. Available at: https://cran.r-project.org/package=ggplot2.

Wieczorek J., Bloom D., Guralnick R., Blum S., Döring M., Giovanni R., Robertson T., & Vieglais D. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*, **7**, e29715.

Wiedenmann J., Baumstark A., Pillen T.L., Meinesz A., & Vogel W. (2001) DNA fingerprints of *Caulerpa taxifolia* provide evidence for the introduction of an aquarium strain into the

Mediterranean Sea and its close relationship to an Australian population. *Marine Biology*, **138**, 229–234.

Wilke C.O. & Wickham H. (2016) *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2."* R package version 0.6.3. Available at: https://cran.r-project.org/package=cowplot.

Williams S.L. & Smith J.E. (2007) A Global Review of the Distribution, Taxonomy, and Impacts of Introduced Seaweeds. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 327–359.

Winter M., Schweiger O., Klotz S., Nentwig W., Andriopoulos P., Arianoutsou M., Basnou C., Delipetrou P., Didziulis V., Hejda M., Hulme P.E., Lambdon P.W., Pergl J., Pysek P., Roy D.B., & Kuhn I. (2009) Plant extinctions and introductions lead to phylogenetic and taxonomic homogenization of the European flora. *Proceedings of the National Academy of Sciences*, **106**, 21721–21725.

Wisz M.S. & Guisan A. (2009) Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC ecology*, **9**, 8.

Womersley H.B.S. (1987) *Handbooks to the Flora of South Australia - Part II.* South Australian Government Printing Division, Adelaide.

Woolley S.N.C., McCallum A.W., Wilson R., O'Hara T.D., & Dunstan P.K. (2013) Fathom out: biogeographical subdivision across the Western Australian continental margin - a multispecies modelling approach. *Diversity and Distributions*, **19**, 1506–1517.

WoRMS Editorial Board (2016) World Register of Marine Species. Available from http://www.marinespecies.org at VLIZ. Accessed 2016-12-20.

Wulff A., Iken K., Quartino M.L., Al-Handal A., Wiencke C., & Clayton M.N. (2009) Biodiversity, biogeography and zonation of marine benthic micro- and macroalgae in the Arctic and Antarctic. *Botanica Marina*, **52**.

Yesson C., Brewer P.W., Sutton T., Caithness N., Pahwa J.S., Burgess M., Gray W.A., White R.J., Jones A.C., Bisby F.A., & Culham A. (2007) How Global Is the Global Biodiversity Information Facility? *PLoS ONE*, **2**, e1124.

Young P.S., Zibrowius H., & Bitar G. (2003) *Verruca stroemia* and *Verruca spengleri* (Crustacea: Cirripedia): distribution in the north-eastern Atlantic and the Mediterranean Sea. *Journal of the Marine Biological Association of the United Kingdom*, **83**, 89–93.

Vander Zanden M.J. & Olden J.D. (2008) A management framework for preventing the secondary spread of aquatic invasive species. *Canadian Journal of Fisheries and Aquatic Sciences*, **65**, 1512–1522.

Zenetos A., Gofas S., Morri C. *et al.* (2012) Alien species in the Mediterranean Sea by 2012. A contribution to the application of European Union's Marine Strategy Framework Directive (MSFD). Part 2. Introduction trends and pathways. *Mediterranean Marine Science*, **13**, 328–352.

Zhang Y. & Grassle J.F. (2002) A portal for the Ocean Biogeographic Information System. *Oceanologica Acta*, **25**, 193–197.

Zheng B. & Agresti a (2000) Summarizing the predictive power of a generalized linear model. *Statistics in medicine*, **19**, 1771–81.

Zhou Z.-H. (2012) *Ensemble Methods: Foundations and Algorithms.* Chapman & Hall/CRC, London, UK.

Zurell D., Jeltsch F., Dormann C.F., & Schröder B. (2009) Static species distribution models in dynamically changing systems: how good can predictions really be? *Ecography*, **32**, 733–744.

Summary

The increased anthropogenic pressure on the marine environment through over-use and overfishing, invasion of species and global climate change has led to an urgent need for more knowledge on the marine ecosystem. Marine species distribution modelling is an important element of marine ecosystem management. It is relied upon by marine spatial planning for i.e. predicting biological resources, the design of marine protected areas, the designation of essential fish habitats, the assessment of species invasion risk, pest control, human-animal conflict prevention, ….

This study aims to improve and contribute to the process and understanding of marine species distribution modelling in order to facilitate an in depth study of the trends, vectors and distribution of introduced seaweeds in Europe. More specifically we wanted to 1) provide quality indicators for the marine species distribution data available in the Ocean Biogeographic Information System (OBIS), 2) make global datasets for species distribution modelling in the past, current and future climate more accessible in R, 3) explore the relevance of different predictors of marine species distributions with MarineSPEED, a marine benchmark dataset of more than 500 species, 4) investigate the introduction history and trends in introduced seaweeds in Europe, 5) evaluate the risk of aquarium trade as a vector for future introductions of seaweeds and 6) study the ability of species distribution modelling to predict the introduction and spread of introduced seaweeds and propose a method for identifying candidate areas for further spreading under climate change.

The first part of this thesis concerns general aspects of marine species distributions, the environmental data used for modelling and the relevance of marine predictors of species distributions.

In **Chapter 2**, different steps are developed to analyse the quality and completeness of the distribution records within the European and international Ocean Biogeographic Information Systems (EurOBIS and OBIS). Records are checked on data format, completeness and validity of information, quality and detail of the used taxonomy and geographic indications and whether or not the record is an outlier. The corresponding quality control (QC) flags not only help users with their data selection, they also help the data management team and the data custodians to identify possible gaps and errors in the submitted data. Through the Biology portal of the European Marine Observation and Data Network (EMODnet Biology), a subset of EurOBIS records—passing a specific combination of these QC steps—is offered to the

users. Through LifeWatch, users can upload their own data and check them against a selection of the quality control procedures. The R package *robis* allows users to query the QC flags of distribution records and additionally filter distribution records based on these.

In **Chapter 3**, we present the open source R package *sdmpredictors*. It allows the end user to download terrestrial and marine environmental layers for the past, current and future climates. *sdmpredictors* contains metadata, statistics and pairwise correlations for the available datasets and layers. These correlations between predictors can be subsequently grouped and plotted. Currently *sdmpredictors* contains data from WorldClim, ENVIREM, Bio-ORACLE and MARSPEC at 5 arcmin resolution and in the Behrmann equal area projection with a resolution of 7 kilometres.

Chapter 4 aims to investigate marine predictor relevance as a function of modelling algorithms and settings for a global dataset of marine species. Additionally, we present the standardized benchmark dataset MarineSPEED and promote its use for methodological SDM studies. For MarineSPEED, we selected well studied and identifiable species from all major marine taxonomic groups. Distribution records were compiled from public sources (e.g. OBIS, GBIF, Reef Life Survey) and linked to environmental data from Bio-ORACLE and MARSPEC. Using this dataset, predictor relevance was analysed under different variations of modelling algorithms, numbers of predictor variables, cross-validation strategies, sampling bias mitigation methods, evaluation methods and ranking methods. SDMs for all combinations of predictors from 8 correlation groups were fitted and ranked, from which the top five predictors were selected as the most relevant. For the creation of the benchmark dataset we collected two million distribution records from 514 species across 18 phyla and made them available with associated environmental data and cross-validation splits through the open source R package *marinespeed* and at **http://marinespeed.org**. Mean sea surface temperature and calcite are respectively the most relevant and irrelevant predictors of marine species distributions. A less clear pattern was derived from the other predictors. The biggest differences in predictor relevance were induced by varying the number of predictors, the modelling algorithm and the sample selection bias correction. Based on the above results we conclude that while temperature is a relevant predictor of global marine species distributions, considerable variation in predictor relevance is linked to the SDM setup. Furthermore, methodological SDM studies should consider the use of a benchmark dataset.

The next three chapters present case studies related to the introduction of seaweeds in Europe.

In **Chapter 5**, we aim to analyse the spatio-temporal trends of introduced seaweeds in Europe. In order to achieve this we assembled a database of seaweed introductions in Europe containing dated observations, origins of the introduced seaweeds combined with an assessments of the uncertainty of the data. Based on this we made a quantitative assessment of the temporal dynamics of primary and secondary introductions, which show that the rate of nonindigenous species being reported for the first time in European waters started declining since the beginning of the 90's. To investigate whether this trend reflects a decline in the number of species being introduced or whether the discovery rate has declined because of factors other than the introduction rate, we analyzed trends in the literature of introduced seaweed species. Contrary to the rate of newly introduced species, the rate of the total number of records remained constant since 1990. The number of papers and authors increased spectacularly from 1970 to 2000 but shows a decrease from then onward. The combination of trends is interpreted as a decline in the rate new species are being introduced. Classifying introduced species according to geographical origin, the decline is mainly attributable to lower numbers of nonindigenous species with a NW Pacific origin being recorded from Europe, while the discovery rates of Lessepsian species or species native to Australasia has remained constant over the years. Given that livestock transfer of shellfish is the principal vector for the introduction of NW Pacific species, it appears that the increased awareness of authorities and stakeholders, and the implementation of policies dedicated to the prevention of introductions, reduce, but not prevent, the introduction of nonindigenous species.

In **Chapter 6** we aim to investigate the potential of aquarium trade as a vector for the introduction of seaweeds in Europe. Firstly, we assessed the seaweed diversity in the European online aquarium retail circuit. This web survey revealed that more than 30 genera are available for online sale into Europe, including known and introduced invasive species. Secondly, we assessed the algal diversity found in local aquaria and on 'live rocks'. As this second approach allowed a direct and accurate identification of the specimens, we targeted not only ornamental species, but also seaweeds that may be accidentally present in the aquarium circuit. By DNA-barcoding we identified no less than 135 species, of which 7 species are flagged as introduced in Europe with 5 of them reported as invasive. Lastly, we build thermal niche models for the current and future climate. These models showed that 23 aquarium species have the potential to thrive in European waters. As expected by the tropical conditions in most aquaria,

southern Atlantic regions of Europe and the Mediterranean are the most vulnerable towards new introductions. From the future climate forecasts, we learn that this risk will increase and shift northwards as global warming proceeds. Overall our data indicates that aquarium trade poses a potential but limited risk of new introductions. However, the large reservoir of macroalgal species in aquaria calls for a cautious approach with the highest risk coming from aquaria in coastal cities and on board of mega yachts.

Chapter 7 focuses on various aspects of modelling the current and future distribution of invasive and introduced seaweeds in Europe. In this study we evaluated the performance of species distribution modelling, trained with native and/or non-European distribution records, as a tool for predicting the spread of invasive seaweeds at various stages of the invasion process. We estimated the level of niche expansion observed under analog and non-analog conditions and assessed which areas in Europe are expected to be disproportionally impacted by migrations of introduced seaweeds due to climate change. Our results indicate that due to considerable niche expansion in non-analog conditions including only native records is generally not sufficient to predict the range of invasive species. Including distribution records from non-European invaded regions on the other hand significantly increases the predictive power of the models and reduces the measured niche expansion in analog and non-analog conditions considerably. Based on forecasts of the distribution of 15 introduced seaweeds in Europe in the future climate, we created European change and turnover maps. These maps predict an increased habitat suitability in northern Europe (northern UK, Scandinavia, Iceland), while southern European regions are likely to become less suitable. In addition to the overall picture, uncertainty in the estimates is apparent for specific regions and this uncertainty correlates only moderately to changes in habitat suitability.

Finally, in **Chapter 8** we highlight various modelling data and uncertainty related aspects of this thesis. Furthermore, we provide some future perspectives with respect to modelling marine species niches and distributions.

Samenvatting

De toenemende blootstelling van het mariene milieu aan een groeiende antropogene druk door overmatig gebruik, overbevissing, de introductie van invasieve soorten en globale klimaatsverandering heeft geleid tot een dringende behoefte aan meer kennis over het mariene ecosysteem. Het modelleren van de verspreiding van mariene soorten is een belangrijk onderdeel van het beheer van mariene ecosystemen. Binnen maritieme ruimtelijke ordening wordt het gebruikt voor het ontwerp van natuurreservaten, voorspellen van biologische bronnen, in kaart brengen van vishabitat, inschatten van het invasieve risico, mens-dier conflictpreventie, …

Deze studie streeft naar het verbeteren en bijdragen aan het proces van en begrip over het modelleren van de verspreiding van mariene soorten om een diepgaande studie van de trends, introductievectoren en verspreiding van geïntroduceerde zeewieren in Europa te vergemakkelijken. Meer specifiek wilden we 1) kwaliteitsindicatoren leveren voor de verspreidingsgegevens van de mariene soorten beschikbaar in het Ocean Biogeographic Information System (OBIS), 2) wereldwijde datasets, geschikt voor het modelleren van de verspreiding van soorten, beschikbaar maken in R, 3) de relevantie van verschillende omgevingsvariabelen voor het modelleren van de verspreiding van mariene soorten onderzoeken met MarineSPEED, een benchmark dataset met meer dan 500 mariene soorten, 4) de geschiedenis en trends in geïntroduceerde zeewieren in Europa te onderzoeken, 5) het risico van aquariumhandel beoordelen als vector voor toekomstige introducties van zeewieren en 6) een haalbaarheidsstudie van het gebruik van verspreidingsmodellen om de introductie en verspreiding van uitheemse zeewieren en om potentiële risicogebieden voor uitheemse zeewieren in het huidige en toekomstige klimaat te voorspellen.

Het eerste deel van dit proefschrift betreft algemene aspecten van de verspreiding van mariene soorten en het gebruik en de relevantie van omgevingsvariabelen gebruikt voor het modelleren van de verspreiding van mariene soorten.

In **Hoofdstuk 2** worden verschillende stappen voorgesteld om de kwaliteit en volledigheid van de verspreidingsinformatie in de Europese en internationale oceanische biogeografische informatiesystemen (EurOBIS en OBIS) te analyseren. De verspreidingsinformatie wordt gecontroleerd op dataformaat, volledigheid en correctheid van informatie, kwaliteit van de gebruikte taxonomische en geografische

aanduidingen en of de registratie al dan niet een statistische uitschieter is. De kwaliteitscontrole (QC) vlaggen helpen niet alleen eindgebruikers met hun data-selectie, maar helpen ook het data management team en de databeheerders om mogelijke fouten in de ingediende gegevens te identificeren. Via het Biology portaal van het European Marine Observation en Data Network (EMODnet Biology) wordt een deel van de EurOBIS-registraties aangeboden aan de eindgebruikers na filtering op basis van een specifieke combinatie van deze QC stappen. Via LifeWatch kunnen eindgebruikers hun eigen gegevens opladen en controleren voor een selectie van de kwaliteitscontrole procedures. Het R pakket *robis* stelt gebruikers in staat om de QC-vlaggen van distributieregistraties van OBIS op te vragen en op basis hiervan te filteren.

In **Hoofdstuk 3** stellen we het R pakket *sdmpredictors* voor. Het biedt eindgebruikers de mogelijkheid om terrestrische en mariene omgevingslagen te downloaden voor zowel het paleoklimaat, het huidige klimaat als het toekomstige klimaat. Bovendien bevat *sdmpredictors* metadata, samenvattende statistieken en correlatiematrices voor de beschikbare lagen. Lagen kunnen op basis van hun correlaties gegroepeerd en gevisualiseerd worden. Momenteel is data beschikbaar van WorldClim, ENVIREM, Bio-ORACLE en MARSPEC met een ruimtelijke resolutie van 5 arcminuten of in de Behrmann equivalente projectie met een resolutie van 7 kilometer.

Hoofdstuk 4 beoogt de relevantie van verschillende mariene omgevingslagen te onderzoeken voor verschillende modelleer methoden en instellingen voor een wereldwijde dataset van mariene soorten. Daarnaast presenteren we de gestandaardiseerde benchmark dataset MarineSPEED om aldus het uitvoeren van methodologische studies omtrent het modelleren van de verspreiding van soorten te faciliteren. MarineSPEED omvat 514 goed bestudeerde en/of gemakkelijk te identificeren soorten behorende tot 18 verschillende *phyla*. In totaal werden twee miljoen verspreiding gegevens uit publieke bronnen (bv. OBIS, GBIF, Reef Life Survey) verzameld en gekoppeld aan omgevingsdata van Bio-ORACLE en MARSPEC. Deze data werd, samen met crossvalidatie datasets van deze data, beschikbaar gemaakt in het R pakket *marinespeed* en op **http://marinespeed.org**. Op basis van ons onderzoek naar de relevantie van de verschillende mariene omgevingsfactoren kunnen we concluderen dat temperatuur de meest relevante en calciet de minst relevante factoren zijn voor het modelleren van de verspreiding van mariene soorten. De grootste variaties in de relevantie van factoren werden veroorzaakt door het gebruik van een verschillend aantal variabelen in de modellen, de gebruikte modelleer methode en het al dan niet gebruik maken van methodes om onevenwichtige rapportering van soorten op te vangen. Met deze studie hebben

we aangetoond dat MarineSPEED een goed instrument is voor het uitvoeren van studies omtrent de methodologie van het modelleren van soorten. Bovendien kunnen we concluderen dat hoewel temperatuur zeer relevant is, er aanzienlijke variatie in de relevantie zichtbaar is wanneer de wijze van modelleren gevarieerd wordt.

De volgende drie hoofdstukken belichten verschillende aspecten van uitheemse zeewieren in Europe.

In **hoofdstuk 5** streven we ernaar om de spatio-temporele trends van uitheemse zeewieren in Europa te analyseren. We hebben hiervoor een Europese database met gedateerde waarnemingen van uitheemse zeewieren, hun oorsprong en een beoordeling van de onzekerheid van de gegevens. Op basis hiervan hebben we de temporale dynamiek van primaire en secundaire introducties geanalyseerd en aangetoond dat het aantal uitheemse zeewiersoorten dat jaarlijks gerapporteerd wordt, sinds het begin van de jaren 90 begon te dalen. Om te achterhalen of deze trend het gevolg is van een daling in de geïntroduceerde soorten of het gevolg van andere factoren, hebben we de temporale dynamiek van het aantal waarnemingen, publicaties en auteurs met betrekking tot uitheemse zeewiersoorten geanalyseerd. In tegenstelling tot het aantal nieuwe uitheemse soorten bleef het aantal waarnemingen sinds 1990 constant. Het aantal publicaties en auteurs daarentegen steeg spectaculair tussen 1970 en 2000, maar daarna nam het aantal publicaties en in mindere mate het aantal auteurs af. De combinatie van trends wordt geïnterpreteerd als een afname van het aantal ingevoerde soorten. Op basis van de waarschijnlijke geografische oorsprong van de uitheemse soorten is deze afname vooral te danken aan een afname van het aantal nieuwe uitheemse soorten afkomstig van het noordwesten van de Stille Oceaan. Aangezien het transport van levende schelpdieren de voornaamste introductie vector is voor soorten uit het noordwesten van de Stille Oceaan, blijkt dat het toenemende bewustzijn van autoriteiten en het gevoerde beleid ter voorkoming van introducties leidt tot een afname maar niet voorkoming van de introductie van nieuwe uitheemse zeewieren in de Europese wateren.

In **hoofdstuk 6** streven we ernaar om het potentieel van aquariumhandel als vector voor de introductie van zeewier in Europa te bepalen. Ten eerste hebben we de zeewier diversiteit in het Europese online aquarium handelscircuit beoordeeld. Hierbij stelden we vast dat meer dan 30 genera vrij online beschikbaar zijn in Europa, waaronder bekende invasieve soorten. Ten tweede hebben we de diversiteit aan soorten in lokale aquaria en op 'levende rotsen' gemeten. Aangezien deze

tweede aanpak de directe en nauwkeurige identificatie van de soorten mogelijk maakte, richtten we ons niet alleen op siersoorten, maar ook op zeewieren die per ongeluk aanwezig zijn in aquaria. Door DNA-barcoding identificeerden we niet minder dan 135 soorten, met 7 uitheemse soorten in Europa, waarvan 5 van hen bekend staan als invasieve soorten. Ten slotte bouwen we thermische niche modellen voor het huidige en toekomstige klimaat. Deze modellen tonen aan dat 23 aquarium soorten het potentieel hebben om in Europese wateren te gedijen. Zoals verwacht door de tropische omstandigheden in de meeste aquaria, zijn de zuidelijke Atlantische regio's van Europa en de Middellandse Zee het meest kwetsbaar voor nieuwe introducties. Dit risico zal in het toekomstige klimaat vergroten en noordwaarts opschuiven met een toenemende opwarming van het klimaat. Algemeen blijkt uit onze gegevens dat aquariumhandel een beperkt risico op nieuwe introducties van uitheemse soorten inhoudt. Het grote aantal soorten zeewiersoorten in aquaria vraagt echter om een voorzichtige aanpak.

Hoofdstuk 7 richt zich op verschillende aspecten van het modelleren van de huidige en toekomstige verdeling van invasieve en uitheemse zeewieren in Europa. In deze studie onderzoeken we in hoeverre modellen van de verspreiding van soorten in staat zijn om de introductie en verspreiding van uitheemse zeewiersoorten in Europa te voorspellen. Deze resultaten linken we vervolgens aan de mate van niche uitbreiding in analoge en niet-analoge klimatologische omstandigheden. Op basis hiervan hebben voor 15 uitheemse zeewiersoorten verspreidingsmodellen gemaakt voor het huidige en toekomstige klimaat om aldus risicogebieden te identificeren. Onze resultaten wijzen erop dat het gebruik van verspreidingsgegevens van het oorsprongsgebied voor het modelleren van een soort leidt tot modellen met een onvoldoende voorspellende capaciteit. Het gebruik van data uit andere uitheemse gebieden en/of Europese data vergroot de voorspellende kracht van de modellen aanzienlijk. De veranderingskaarten, gemaakt op basis van modellen van 15 uitheemse zeewieren, voorspellen voor het klimaat in het jaar 2100 een verhoogde habitat geschiktheid in Noord-Europa (Noord-Engeland, Scandinavië, IJsland), terwijl Zuid-Europese en vooral mediterrane regio's waarschijnlijk minder geschikt worden. Naast het algemene beeld is de onzekerheid in de schattingen duidelijk voor specifieke regio’s en staan deze los van de mate van verandering.

Ten slotte belichten we in **hoofdstuk 8** diverse aspecten met betrekking tot de gebruikte data en de onzekerheid van de modellen in dit proefschrift. Daarnaast bieden we toekomstige perspectieven met betrekking tot het modelleren van de niche en verspreiding van mariene soorten.

Acknowledgements

Olivier, thank you for everything you did, without your trust, money, support with my research and many more small and big things no one would be reading this.

Members of the examination committee, thank you for reviewing my work and providing helpful suggestions.

Ghent University, thank you for the (computing) infrastructure, the libraries and the excellent doctoral training and master courses.

INVASIVES, my first working day was the 1st project meeting, I really enjoyed the three meetings and really appreciate having been a part of this project.

VLIZ and especially the VMDC team, thank you both for the funding, excellent infrastructural support and help during my work on this thesis, but also thank you for being such pleasant colleagues.

OBIS, thank you for the excellent collaboration and workshop.

Current and former colleagues at Phycology and PAE, we didn't collaborate a lot but I really enjoyed the atmosphere and the time spend together.

Students under my supervision, thank you for your positive attitude, effort and contribution to this thesis.

Vrienden, bedankt om te zijn wie jullie zijn, of het nu voor een babbeltje is of voor een feestje.

Familie, grootouders, ouders, broers, biologisch of niet, bedankt voor alle fijne momenten samen, bedankt voor jullie invloed op mijn leven, merci pour tous les beaux moments.

Jasper en Lander, bedankt lieve kleine, of niet meer zo kleine, schattige uitbundige jongetjes om samen spelenderwijs het leven te ontdekken. Het boekje is nu klaar, maar ik ga het niet voorlezen.

Liefste, bedankt om samen met mij over berg en dal te wandelen!